

Avoid Wacky (Genome) Races

As the final polish is given to two versions of the human genome sequence, tensions between the competing camps are inevitably high. But excited banter could be damaging to both if it gets out of hand.

On Monday of this week, the article reporting the complete sequence of human chromosome 21 was published on *Nature's* website (the paper version will be published in next week's *Nature*; **405**, 311–319; 2000). The completion of the second human chromosome — the first, chromosome 22, was published last December (see *Nature* **402**, 489–495; 1999) — is a milestone in genomics, and a notable achievement for the Japanese and German groups responsible. It also presages a long, hot summer for genomics, from which we will emerge with at least one, and probably two, versions of the human genome sequence.

The international, publicly funded Human Genome Project (HGP) consortium, to which both the chromosome 21 and 22 teams belong, is likely to finish its draft sequence within weeks. This is generally expected to be 90 per cent complete at 4 × coverage — meaning that each bit of the genome has been sequenced four times to improve accuracy. Meanwhile, the private company Celera Genomics is in the closing stages of 'shotgun sequencing' and starting to assemble the DNA fragments that have been sequenced. Exactly how long it will take Celera to produce an ordered sequence is not clear. But, like the HGP sequences, it will certainly be within the next few months.

How are we to compare these parallel projects? The view of Harold Varmus, director of the National Institutes of Health, that deciphering the sequence of the human genome should be seen less as a 'road race' and more as the discovery of a new continent, "with competing enterprises exploring different areas of the terrain, using somewhat different technologies", certainly has merit. And Eric Lander's ad-

monition that "the whole finish-line mentality is silly" (see page 111) is well taken. But the temptation, both for many scientists involved in the respective projects, and for many journalists trying to report on their activities, to describe the situation as a road race is irresistible. And, up to a point, it is true.

How the relationship between the two groups will be characterized in the coming weeks depends heavily on the behaviour of those involved. So far, good sense and manners have only occasionally been overtaken by bouts of claim and counter-claim, hype and counter-hype. Then, the most appropriate road-race image was that of a scientific *Wacky Races* (the cult 'sixties cartoon series and current kids' computer-game hit; readers bemused by the analogy should consult <http://www.hotink.com/wacky/>). Some coverage already suggests that the genome community has its Professor Pat Pending, a Gruesome Twosome, an Ant Hill Mob, even a Dick Dastardly.

Good knockabout stuff, perhaps, and no serious damage has been done, at least not so far. But under the intense glare of public, media, political and business interest, the protagonists may be tempted into a more damaging public slanging-match. This would be even more irresistible to the media — and bring altogether more damaging analogies, perhaps even threatening the continued financial support that both the public and private efforts will require. This shouldn't be allowed to happen. Much is at stake over the summer, not least the reputation of science as one of the few domains whose ultimate goal, for all its internal intrigues and personal competitiveness, remains the common good. ■

A global solution to global problems?

Time will judge a body being set up this week to address the scientific aspects of topics of international concern.

It has become commonplace over the past century to remark on the need of modern governments for access to high-level scientific advice. Various institutional forms have been developed through which this advice can be channelled. One of the most successful has been the National Research Council, the operating arm of the National Academy of Sciences (NAS), which originated in the US government's need for guidance on military technologies during the First World War. Now the emergence of a set of trends and issues identified with 'globalization' — ranging from genetically modified foods to the common challenges of science education — has triggered an attempt to reproduce this success at the international level.

This week, representatives from 80 academies of science around the world are meeting in Tokyo to endorse the creation of an InterAcademy Council. The purpose of this body, according to the committee that is drafting a proposed constitution, would be "to facilitate the provision of advice and recommendation on issues of global importance for those organizations and governments formally requesting such an input". Bruce Alberts, the president of the NAS, a prime instigator of the move, has suggested the need to address a fundamental question: "How can the world's scientists ensure that rationality, rather than misinformation, forms the basis for decision-making in this ever more confusing world?"

Anyone who doubts the pertinence of such a question need look no further than South Africa, currently embroiled in a heated debate over an issue that many scientists regard as long since resolved — whether there is a causal relationship between HIV and AIDS (see page 105). But there are also good reasons for scepticism. Do international bodies such as United Nations agencies operate in such a way that 'science-based' advice — however impartial — will be of practical use to them? Do sponsors exist ready to commit the type of funds that will be required for in-depth studies of controversial topics? Can an international body that works through consensus operate with the speed and flexibility necessary to address urgent issues on the global agenda? Can duplication be avoided with other bodies, such as the International Council for Science (ICSU)? Will the strong US influence undermine claims to regional balance? And, perhaps most importantly, do the institutions of science retain sufficient authority in political circles to make the operation worthwhile?

There are no simple answers. Those responsible admit that they face major hurdles, but express optimism, based partly on enthusiasm from parts of the scientific community and partly on informal soundings with potential sponsors. Others, while harbouring doubts, are ready to go along with the project in the hope that even partial success will be worthwhile. Time will deliver the ultimate verdict. ■





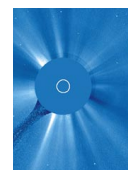
Cold feet
Greens revolt over
German fusion
research
p107



Slow progress
Canadian neutron
scattering facility
faces funding crisis
p107



Net gains
US efforts to boost
public confidence in
GM food via the web
p108



All lined up
The planets take
their positions in
the celestial queue
p111

South Africa turns to research in the hope of settling AIDS policy...

Pretoria

The relationship between HIV and AIDS is set to be put to the test in South Africa in a bid to defuse the controversy surrounding comments made by President Thabo Mbeki. He has often stated that he maintains an open mind on whether HIV is the cause of AIDS (see *Nature* 404, 911; 2000).

The inaugural meeting of an international panel, set up by Mbeki to advise on South African AIDS policy, announced the proposals last weekend. South Africa's Medical Research Council (MRC) will team up with the Centres for Disease Control (CDC) in Atlanta, Georgia, together with two prominent AIDS 'dissidents', who dispute that HIV causes AIDS, to devise a series of surveys to investigate the relationship between the disease and the virus.

According to MRC president Malegapuru Makgoba, the surveys could involve the clinical identification of a sample of AIDS sufferers, who would be tested for HIV. Another possibility is an epidemiological study correlating HIV-positive children with the HIV status of their parents.

A task force made up of Makgoba, Helene Gayle of the CDC — which has placed its database at the team's disposal — and two prominent dissidents, Berkeley biochemist Peter Duesberg and biotechnologist Harvey Bialy, has been appointed to conceptualize exactly what research needs to be done. Makgoba told *Nature* that the establishment of a national register for HIV/AIDS in South Africa — in which the CDC could presumably assist — is also on the cards.

Khotso Mokhele, president of the National Research Foundation and one of three facilitators on the international panel, said after the meeting that funding would be sought from the South African government and other bodies once the agenda had been clarified. He emphasized that existing knowledge, based on completed studies, should provide the basis for any new work.

In opening the two-day meeting, Mbeki emphasized the high level of heterosexual AIDS transmission in South Africa. He referred to the first South African paper on AIDS, published in 1985, which predicted



Shifting position?
Mbeki is showing signs of having second thoughts about his 'open-minded' stance on the link between HIV and AIDS.

that the disease would remain largely confined to male homosexuals. The panel's task would be to try to explain why this situation had not changed in the West, but had in South Africa, he said, adding that the conclusions would have a direct bearing on the government's response to the problem.

But the composition of the 36-member panel is itself controversial, as just under half of its members are dissidents. "The panel

has pretty well everyone on it who believes that HIV is not the cause of AIDS, and about 0.0001 per cent of those who oppose this view," comments AIDS researcher John Moore of Weill Cornell Medical College.

Announcing the panel last week, health minister Manto Tshabalala-Msimang said that it would consider the causes of the immune deficiency leading to AIDS, and the best response to the pandemic in a local con-

...as editor defends publishing key AZT paper

London

A peer-reviewed paper at the centre of the South African row over the efficacy of the anti-AIDS drug AZT was published as much to stimulate debate as to endorse its conclusions, according to the editor of the journal in which it appeared.

The paper, "A Critical Analysis of the Pharmacology of AZT and its Use in Aids", was published last year as a supplement to *Current Medical Research and Opinion*. Many of the paper's main authors are identified with the 'dissident' research community

that claims that HIV is not the cause of AIDS.

South African president Thabo Mbeki has referred to this paper, which he came across on the Internet, in support of his reluctance to endorse the use of AZT — for example, to prevent the transmission of HIV from pregnant mothers to their babies.

But Peter Clarke, the journal's managing editor — who at the time was also in charge of editorial decisions — describes the paper as "more of a review rather than the result of original

research *per se*". The paper was refereed over an "extended period of time" and views "varied widely", says Clarke.

"I felt that what the authors were trying to do — that is, critically appraise existing published material — was legitimate and deserved an opportunity to be made public." Clarke adds that he took many months to reach a decision: "At the end of the day I took the value judgement that it needed to go into the public domain to be debated."

Natasha Loder

♦ <http://www.librapharm.co.uk/cmro>

text. It would also investigate why HIV/AIDS was heterosexually transmitted in southern Africa, and assess drug-based responses, including strategies to prevent mother-to-child transmission, she said.

Panel members include the French discoverer of the AIDS virus, Luc Montagnier, and Clifford Lane of the US National Institutes of Health. Twelve of the panel are US-based, and ten are African, including seven South Africans and representatives from Uganda, Malawi and Senegal. There are also members from Cuba, Mexico and India.

Significantly, with the exception of Sam Mhlongo of the Medical University of South Africa, none of the African representatives belong to the dissident camp. But several prominent South African AIDS researchers, all of whom have been outspokenly critical of the dissident movement, were not included. These include Jerry Coovadia of Natal University (convenor of the World AIDS Congress to be held in Durban in July), James McIntyre of the University of the Witwatersrand, immunologist Johnny Sachs, Gary Maartens of the University of Cape Town, and epidemiologist Bryan Williams of the Council for Scientific and Industrial Research.

"Although it's very important to determine innovative strategies for combating AIDS, engaging with fringe groups is not the way forward," says Glenda Gray, director of



Parents and their children could be HIV-tested.

the perinatal HIV research unit at Johannesburg's Baragwanath hospital, who was also excluded from the panel. "These people should never have been given a platform."

Although Tshabalala-Msimang described the meeting as a "wonderful experience", it is said to have been very acrimonious, with the dissidents finding themselves in a minority in each of three groups appointed to discuss the causes, prevention and treatment of AIDS.

Instead of a final round-table discussion, the meeting is understood to have divided into groups representing the orthodox and dissident views.

The panel's chief facilitator, lawyer Stephen Owen of the Institute for Dispute Resolution at the University of Victoria in Canada, said at the press conference follow-

ing the meeting that reaching consensus had not been the objective. "Divergent points of view remain, in very stark terms," he said.

The panel will enter into a "closed Internet debate" over the next four to six weeks, before reconvening in South Africa for a four-day discussion before the start of the World AIDS Congress on 9 July.

The task force is widely interpreted as a face-saving device for Mbeki, who admitted at the meeting's opening that he was "embarrassed to say" that he had "discovered that there had been a controversy about this for some time". Quoting the Irish poet Patrick Henry Pearse, Mbeki pondered whether his having raised the issue was "folly or grace".

There has been some speculation that a last-minute deal to add three extra non-dissident members to the panel — thereby creating an overall majority of non-dissidents — was brokered at a high level between the South African and US governments.

The three additional names had not appeared on the initial list announced last week; neither Tshabalala-Msimang nor Essop Pahad, cabinet minister in the President's office, were prepared to confirm this at the press conference.

Some South African AIDS researchers feel that if the proposed surveys counter the claims of the dissident movement, they will have done a useful service.

Michael Cherry

French research minister targets IT and biotech

Paris

France's new research minister, Roger-Gérard Schwartzberg, promised last week that the government will substantially increase its support for research in information technology and biotechnology.

Schwartzberg says that over four years he will double the research budget of France's computing research agency, INRIA, which received FF408 million (US\$56 million) in state support in 1998. In addition, a computer science division will be created at the Centre National de la Recherche Scientifique.

Schwartzberg plans to use the research ministry's own funds to pay for these increases. He says that the money allocated for new technologies from the National Science Funds and the Funds for Technology Research, which the ministry directs, will increase by 50 per cent next year.

The budget increases were announced by Schwartzberg during his first public speech on his plans for the post. He was appointed to head the research ministry in March after his predecessor, Claude Allègre, was sacked by the prime minister, Lionel Jospin, in a cabinet reshuffle.

Unlike Allègre, who ran a superministry

overseeing higher education and research, Schwartzberg will "have all of his time to dedicate to raising the research budget", says Vincent Courtillot, who served as research director under Allègre and will retain a similar position in the new ministry.

One priority highlighted by Schwartzberg last week is to boost spending for life-sciences research. He also plans to set aside FF200 million as seed money for biotechnology companies, as well as a national network of 'incubators' specialized in raising biotech start-ups.

On most fronts, Schwartzberg will continue with the policies of his predecessor. But he has broken with Allègre's opposition to building a new synchrotron facility on French soil (see *Nature* 404, 533; 2000). The new minister, while continuing a partnership with the British government and the Wellcome Trust to build the synchrotron Diamond in Britain at the Rutherford Appleton Laboratory near Oxford, is also pursuing a third-generation machine in France.

Last week Schwartzberg met with British science minister Lord Sainsbury and discussed the possibility of British cooperation in the French project, Soleil. A



Schwartzberg: wants to help start-ups.

statement issued after the meeting said that the two facilities "would enable the scientific community in the future to cover a wide range of applications of synchrotron radiation".

Schwartzberg's appointment coincides with France assuming the presidency of the European Union in July. In last week's presentation, he said he favoured accelerating the creation of a European patent, creating incubators and seed money for new technology companies, and harmonizing the tax system to favour the creation of new companies.

Heather McCabe

German Greens go cold on nuclear fusion

Munich

The Green Party, the junior partner in Germany's ruling coalition, is urging the government to slam the brakes on research into nuclear fusion. It wants Germany to cut its commitments to the national fusion research programme, and to withdraw support for the planned International Thermonuclear Experimental Reactor (ITER).

The Green Party's strategy paper on energy research, published last month, recommends that nuclear research be restricted to areas such as reactor safety. It says that money earmarked for research on new nuclear fusion and fission technologies should be used to increase the efficiency of technologies for renewable energy sources, such as wind and solar energy.

The German fusion community is reacting nervously to such suggestions, aware that energy policy is a core issue for the Greens, and that fusion research has still to demonstrate its feasibility. The Greens have already achieved the government's commitment to one of their major goals — phasing out atomic energy in Germany by 2019 — so the party's position on other energy-related issues has obvious political weight.

"Energy from nuclear fusion will only be available in half a century, if at all," says Hans-Josef Fell, the Greens' parliamentary spokesman on research, and a member of the parliamentary committee on science and technology. "Given the urgency of replacing fossil and atomic energy by new climate-friendly and safe energy sources, it would be fatal to put trust in such a vague and distant option," he says.

ITER, which was designed to study the physics of burning plasma and the engineering problems related to a future power-generating fusion reactor, is a proposed 3.5 billion euro (US\$3.1 billion) collaborative project between Europe, Russia and Japan. Its design outline was slimmed down considerably last year, after the United States, concerned about the project's high costs and scientific uncertainty, withdrew support (see *Nature* 402, 570; 1999).

The Greens would like Germany to follow suit. They want the government to urge the European Commission not to co-fund ITER in its sixth Framework programme of research (FP6), due to start in 2003. Discussions about FP6 will begin this summer.

But the German government has given no signs that its support for fusion research is waning. The research minister, Edelgard Bulmahn, and her state secretary for research, Wolf-Michael Catenhusen, are known for their pragmatic views on scientific projects. It appears unlikely that Bulmahn will take the lead in any possible European opposition to ITER when the European



Uncertain future: the Wendelstein 7-X fusion reactor would be redesigned under Green plans.

science ministers meet next year to decide about future support for the project.

Fusion scientists in Germany hope that pragmatism will also thwart another of the Greens' proposals: a major redesign of the Wendelstein-7X, an experimental fusion reactor being built in Greifswald in eastern Germany by the Max Planck Institute for Plasma Physics (IPP).

The 'stellarator' reactor, which cost DM300 million (US\$135 million), is eastern Germany's largest research facility. It is designed for a plasma-physics experiment using superconducting coils to produce the magnetic field necessary for sustained burn

— an unsolved problem with regard to a functioning fusion power plant.

The Greens say the reactor should be built as planned, but that it should be used almost exclusively for research in conventional low-temperature plasma physics and on a few possible spin-offs from nuclear fusion research.

Alexander Bradshaw, scientific director of the IPP, calls the idea "scientifically and economically absurd, and technically impossible". Bradshaw says he is particularly disappointed because he is sympathetic to Green policies in areas such as climate protection, and appreciates their generally positive attitude towards basic research.

"The Greens can prosper without hitting out at fusion research," says Bradshaw. He adds that the US decision on ITER is irrelevant, as overall US spending on fusion technologies is increasing (see *Nature* 400, 394; 1999).

Bradshaw is confident that the government will base its energy research policy on proposals made last year by the Wissenschaftsrat, Germany's influential science council, rather than by the Greens. The Wissenschaftsrat recommended that all energy options should be left open, and called for a 30 per cent increase in energy research budgets (see *Nature* 397, 375, 1999).

The Greens' proposals will be discussed further in the federal parliament's committee on research.

Quirin Schiermeier

Canada's plans for neutrons stall

Washington

Canada's leading role in neutron scattering is under threat, according to Canadian physicists. An ambitious plan to construct a world-class neutron-scattering facility at Chalk River, Ontario, could collapse unless funds are forthcoming from the government this year, they claim.

Government ministers have approved construction of the Can\$400 million (US\$270 million) facility in principle. But money to start the project was not included in this year's budget when it was announced in February (see *Nature* 404, 8; 2000).

"We don't understand why it didn't appear" in the budget, says Bill Buyers, a senior scientist on the twenty-strong team behind the proposal.

Now project backers are hoping that the government will announce support for the facility this summer, possibly as part of a 'mini-budget' in the run-up to the next general election. "I'm very optimistic that it will get funded," says Buyers.

A replacement for the 43-year-old National Reactor Universal (NRU), the proposed Canadian Neutron Facility (CNF) would have eleven experimental beamlines and space to add another twelve. It would be built on the same Chalk River site as the NRU, which is expected to close when its operating licence expires in December 2005.

Scientists had hoped that the six-year construction of the CNF would start last year, avoiding a 'neutron gap' when the old reactor shuts down. Canada is already building two medical isotope reactors to replace the NRU, which is the largest source of medical isotopes in the world, accounting for 70 per cent of the total market.

The CNF would be built jointly by Atomic Energy of Canada Limited (AECL) and the National Research Council of Canada. The reactor would serve both as a source of neutrons for scientific research and as a testbed for fuel rods and other nuclear reactor components required by AECL, whose CANDU (Canadian deuterium



Waiting game: plans for the Canadian Neutron Facility still need government funds.

► uranium) designs for nuclear power stations have had some export success.

The double use of the facility means that its fate rests not just on the scientific case for a neutron facility, but also on the perceived importance of AECL's power reactor development programme at a time when few nuclear power stations are being built anywhere in the world.

"The dual use thing is a double-edged sword," says Thom Mason, a Canadian who is leading the project to build a \$1.3 billion neutron source at the Oak Ridge National Laboratory in the United States. "The case for neutron scattering is accepted, but the decision hinges on the future of CANDU, and that is totally decoupled from the scientific community's need for neutrons."

Mason is a strong supporter of the Canadian project. But if lack of funding causes the Canadian team to disintegrate, progress at the US facility, together with a new target chamber at ISIS in the United Kingdom and an imminent decision to build a neutron source in Australia, could all attract people away from Canada.

The 300 or so Canadian researchers who use neutrons say that failure to fund the facility will cause Canada to lose its long-standing leadership in neutron scattering, for which Canadian physicist Bertram Brockhouse won a Nobel Prize in 1994. "Neutron sources in the United States are heavily oversubscribed — we can't get beamtime there," says Buyers.

Researchers see the CNF as completing an important trio of national scientific facilities, together with Triumph, a particle accelerator in British Columbia, and the Canadian Light Source, a synchrotron light source being built at Saskatoon in the province of Saskatchewan.

But the CNF hasn't received financial support from the province of Ontario. And, ironically, the fact that the province is overwhelmingly loyal to the Liberal party of Jean Chretien, the Canadian prime minister, may actually reduce pressure on him to fund construction of the facility.

Colin Macilwain

US reforms rules for telling public about GM food

Washington

The Clinton administration has announced a series of regulatory changes and research proposals intended to shore up public confidence in the government's supervision of genetically modified (GM) food.

The Food and Drug Administration (FDA) will in future require companies that wish to introduce any new transgenic food to provide notice and supporting scientific research 120 days in advance. This information will then be placed on the Internet for public inspection. At present, companies submit this information on a voluntary basis, and it is not automatically made available to the public.

The FDA will also develop guidelines for the voluntary labelling of GM food, and will permit producers of food containing no GM organisms to label it as such.

But the changes, which were announced last week by the White House in conjunction with the FDA, the US Department of Agriculture (USDA) and the Environmental Protection Agency (EPA), did not include any mandatory requirement for the labelling of GM foods.

Although widely expected, this omission led most environmental groups to reject the changes, which they branded as cosmetic. Farming and industry groups, meanwhile, warmly welcomed the announcement. The National Corn Growers Association, which represents most large maize farmers, said the changes matched its own policy and position, as stated during public hearings

conducted by the FDA last autumn (see *Nature* **402**, 571; 2000).

Farmers and the agricultural biotechnology industry have been pressing the government to make such changes, in the hope that they will strengthen US public confidence in GM food. GM foods are already ubiquitous in the food chain in the United States, where around half of this year's soybean crop and one-third of the maize will be transgenic.

The industry is concerned that European rejection of the technology will spill over into the United States, where GM crops were introduced after extensive scientific review but minimal public debate.

The USDA, FDA and EPA also pledged to coordinate their research programmes on the safety and risk assessment of agricultural biotechnology, although the amount of additional money to be made available for this was not specified. The USDA also said it will create standards to certify the various testing procedures that are available to establish whether foods contain GM organisms.

The White House Office of Science and Technology Policy and the Council of Environmental Quality said that they would conduct a further six-month study on the regulation of agricultural biotechnology. Under a long-standing arrangement, the FDA, EPA and USDA share responsibility for this regulation, depending on the intended function of the genetic modification.

Colin Macilwain

Italian genomics boost retained

Munich

In Italy, scientific initiatives tend to change with governments. But prime minister Giuliano Amato's new administration has retained a plan to launch strategic research programmes in human genomics and neuroscience.

Only days before the previous government fell last month, it issued a decree setting up two national committees to define the strategies. The decree was the brainchild of Vincenzo Sica, undersecretary of state for research, who has been reappointed to his position.

The committees, which met for the first time earlier this week, will put forward both scientific and financial proposals. If all goes according to Sica's plan, these will be worked into next year's research budget.

Sica hopes that substantial sums of money will start to flow early next year. "We'll need at least IL100 billion [US\$46 million] per year, and probably a lot more," he says.

Given the unpredictability of Italian politics — and the infighting at the ministry over the control of any genomics programme — this timetable may be a little optimistic. But if it works it will give Italian biology, long hampered by underfunding and poor management, a shot in the arm.

"Like all areas of biomedicine, the small amounts of funding available to neuroscience research are distributed thinly across a large number of research groups, and not strategically coordinated," says Piergiorgio Strata, professor of

► neurophysiology at the University of Turin, who will coordinate the neurosciences committee.

Genome research has struggled to gain a toe-hold in Italy. "Italy has been the only one of the G8 countries not to have a genome project, or any serious money for genome research from the state," says Andrea Ballabio, director of the Telethon Institute of Genetics and Medicine in Milan, and a member of the genomics committee, now called the Committee for Medical Genetics.

This concern led Ballabio, along with several other top geneticists and molecular biologists in Italy, to create an informal group called the Genome Task Force. This spent more than a year working in secret — to escape external pressures — on a strategy paper for genome research. It had planned to present the paper to the research ministry, when it was overtaken by Sica's initiative.

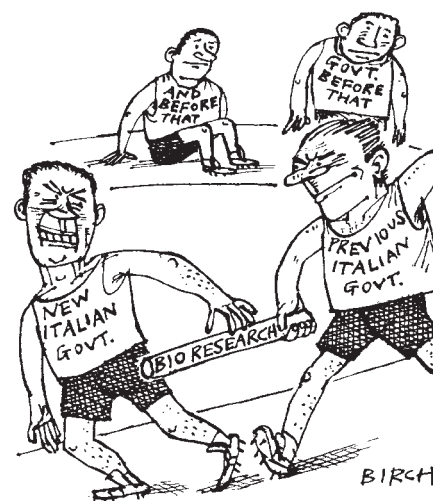
The paper suggests the creation of national centres to service fields such as genotyping, bioinformatics, expression

profiling and proteomics. It also advocates the support of projects to develop animal models for genetic diseases, and to build on strengths in population genetics. This paper is likely to be a springboard for the national committee's discussions.

"The committee is an important first step for us," says Ballabio, "and it is certainly not too late for Italy to catch up — the interesting genome work is only just starting now that the sequences are being completed."

Although the genomics and neurosciences national committees will define broad strategic goals, they will not control the distribution of grant money. This will be based on the advice of a panel of international experts.

Sica, a neuropathologist from the University of Naples, was appointed undersecretary of state for research last December. Knowing that his time in office would be short — government collapses notwithstanding, general elections are due in Italy next year — he has concentrated on



pushing through initiatives in what he believes to be the two most important research areas. Many of the scientists on the two committees feel that Sica's initiative would almost certainly have evaporated had he not been reappointed.

Alison Abbott

Israel debates raising commitment to CERN

Jerusalem

Keen to boost its role in the European scientific community, Israel's science ministry is considering whether to apply for full membership of the European Laboratory for Particle Physics (CERN). But although some Israeli physicists are enthusiastic, others worry about the financial and scientific implications.

The 1992 agreement granting Israel observer status, similar to other countries such as the United States, Japan and Russia, is about to expire. Under this agreement, Israel pays 20 per cent of full membership dues — about 2 million Swiss francs (US\$1.2 million) a year. In exchange, Israeli researchers and postdocs can participate in CERN research programmes, and Israeli companies can bid for CERN tenders.

Renewal of observer status seems the most likely short-term outcome. But some Israeli particle physicists are pushing hard for full membership. The reasons, they admit, are as much political as scientific.

"Israel is no longer a Third World country. It has to take responsibility for the things that large countries do. It has to support first-class science," says Giora Mikenberg, professor of particle physics at the Weizmann Institute of Science.

Mikenberg, who leads the muon project at CERN's Large Hadron Collider, thinks that membership of CERN could lead to Israel becoming a full partner in other European scientific programmes — such as the European Space Agency — and



Collision course: Israeli physicists are divided over the benefits of full membership of CERN.

would show that Israel was a scientific power of the first rank. Full membership would also allow Israelis to be CERN employees and to participate in scientific decision-making there.

The main problem, Mikenberg acknowledges, is the price tag. The ministries of science and industry, which fund Israel's

observer status at CERN, will not be able to finance full membership, meaning that the cabinet would have to allocate money from the national treasury, he says.

Israeli minister of science Matan Vilnai visited CERN recently and was impressed by what he saw. An informal ministry forum of industrialists, researchers and government officials has discussed full membership and made a favourable recommendation to Vilnai.

Sources in the ministry claim that Vilnai discussed the possibility of full membership with CERN officials. But CERN director-general Luciano Maiani denies having any conversations about the matter with Vilnai or anyone else. "The issue is not on the agenda," he insists. "There is no official move in that direction."

Other Israeli physicists are doubtful about the scientific benefits of full membership, and worry that it might actually have negative consequences. Shmuel Nussinov, a theoretical physicist at Tel Aviv University, says that he would welcome closer ties with CERN, but worries that the additional funding needed might come at the expense of existing research.

Another physicist says that, at present, Israeli physicists can participate in CERN research projects but have other options. Full membership of CERN could mean putting all the country's eggs in one basket, he says, and might make it harder for Israelis to find funding to take part in particle-physics research elsewhere.

Haim Watzman

Troubled US fusion project receives cash injection

Washington

The US Department of Energy has agreed to provide more funds for the building of the National Ignition Facility (NIF) at the Lawrence Livermore National Laboratory in California. It will contribute an extra \$200 million–\$250 million over the next two years, and will extend the construction schedule by four years, to 2008.

Confirming the worst predictions of the project's critics (see *Nature* 403, 469; 2000), Thomas Gioconda, assistant secretary of energy for defence programmes, said last week that the cost of the troubled project is now expected to exceed the original estimate of \$1.2 billion by between \$750 million and \$1 billion.

The NIF is the world's largest laser project, and aims to cause fusion by firing 192 huge lasers at a tiny deuterium target. The project is central to the US government's stockpile stewardship programme, which will maintain nuclear weapons — and the army of scientists and engineers that work with them — in the absence of nuclear testing.

Just how much the NIF is over budget will be confirmed in the next few weeks, when the General Accounting Office publishes a report on the project. But the scale of the cost overrun has shocked staff and members of Congress, where the Department of Energy's plan to complete the NIF must win approval.

Gioconda said that the money to complete the programme will come from the rest of the nuclear weapons research budget, and "predominantly" from the Livermore laboratory's \$1 billion annual budget. Previously, Bill Richardson, the energy secretary, had said that it would all come from Livermore's other programmes.

Under the new decision, announced on 3 May, a first phase of the NIF, with 96 working lasers, will be completed in 2004, with the rest becoming operational by 2008. The facility would aim to reach ignition — the point at which it produces a self-sustaining fusion reaction — in 2011.

Richardson's choice falls short of an alternative favoured by the Livermore laboratory, which would inject an extra \$390 million over the next two years to bring the project back on schedule. The laboratory says that it will now support Richardson's plan.

Despite the extraordinary scale of the cost overruns at the NIF, the indications are that Washington will continue to support the project, say administration and congressional staff.

Colin Macilwain

Medical tests cost Lawrence Berkeley \$2.2 million

San Francisco

In the first class-action suit claiming discrimination and invasion of privacy related to genetic and other medical testing, the Lawrence Berkeley National Laboratory in California has reached a provisional \$2.2 million settlement with a group of employees. But the laboratory continues to deny any wrong-doing.

Seven former and current employees filed the suit in 1995, arguing that, by testing its workers for pregnancy, sickle-cell trait and syphilis in occupational medical examinations without their knowledge, the US Department of Energy (DoE) laboratory was guilty of sexual and racial discrimination and invasion of privacy. The employees said the tests were especially harmful because they dealt with intimate matters, including genetic information.

In June 1996, the US District Court for the Northern District of California granted a petition by the laboratory to have the case dismissed. But that decision was reversed in 1998 by the US Court of Appeals (see *Nature* 393, 611; 1998).

Judge Stephen Reinhardt, who wrote the opinion, noted that "the conditions tested for were aspects of one's health in which one enjoys the highest expectations of privacy". He pointed out, for example, that a test for sickle-cell trait — one copy of the recessive gene for sickle-cell anaemia — could reveal information about family history and reproductive decision-making.

By providing genetic data, the test could disclose paternity, as well as the potential health of future generations. The laboratory said the tests were part of routine and comprehensive medical screening to which the employees had consented, and that they were consistent with good medical practice.

Lawrence Berkeley laboratory has agreed to settle the case, while still denying any damage to its employees. After about 18 months of mediation, the laboratory said it would pay the seven named plaintiffs \$25,000 each.

The deal set aside another \$100,000 for ten other early claimants and \$440,000 to pay the plaintiffs' legal costs. Awards of up to \$2,000 each would then go to other employees who came forward and said that they were also tested, depending on the tests performed.

Notices went out last week detailing the agreement to former and current employees. The court said that potential class members included all full- or part-time workers at the laboratory from January 1981 to April 1993.

The group also includes women employees between March 1972 and December 1994, who were probably tested for pregnancy, and African Americans who worked there from March 1972 to June 1995, who were likely to have been checked for sickle-cell trait.

No one involved in the case will comment on whether such testing was carried out at other DoE laboratories. The litigants have agreed not to discuss the case publicly until the court decides in July whether to approve the settlement.

Sally Lehrman

Australia improves its AIMS

Sydney

The Australian Institute of Marine Science (AIMS), based at Cape Ferguson in northern Queensland, is to be upgraded, following a glowing review of the institute's work by Robin Batterham, Australia's chief scientist.

AIMS had been under threat of effective takeover by the country's largest national research agency, the Commonwealth Scientific and Industrial Research Organisation (CSIRO). The CSIRO had proposed that the funds be put towards a new "Marine Research Centre of Excellence" in Townsville, 50 km away, with AIMS being closed and its research absorbed into the centre.

But Batterham said that AIMS should stay put, and be allowed to spend funds on laboratory expansion and a new research vessel. Nick Minchin, the minister responsible for both agencies, has agreed. Last week



Sea view: designed by AIMS, the WetPC lets divers do computerized surveys underwater.

he released A\$17 million (US\$10 million) in funding and commissioned Batterham to undertake a broader review of tropical marine research in Australia.

Peter Pockley

Human Genome Project ends 'rough draft' phase

Washington The Human Genome Project this week announced that it is wrapping up its 'rough draft' of the human genome. The consortium of 16 genome centres has deposited around 85 per cent of the genome with GenBank, the public genome database. All the clones required to claim a rough draft are in the sequencing pipeline, says Eric Lander, director of the Whitehead Institute in Cambridge, Massachusetts. He expects that the remaining 5 per cent necessary to claim rough-draft status will be sequenced soon and will reach the database over the next six weeks.

The rough draft will still contain many gaps, which the project will fill by the end of 2003. Lander applauds the progress the project has made, but decries the way public and private efforts have been depicted as a race. "The whole finish-line mentality is silly," he says. Celera Genomics last month announced that it had completed, but not assembled, a proprietary version of the genome (see *Nature* 404, 691–692; 2000).

Anderson quits Oxford for Imperial College

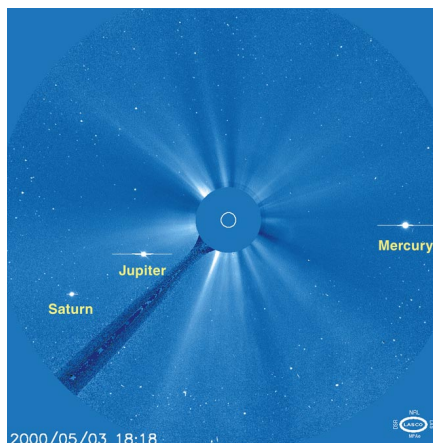
London Epidemiologist Roy Anderson has announced that he intends to resign his position as the Linacre Professor of Zoology at the University of Oxford. He will take a post in the school of medicine at Imperial College of Science, Technology and Medicine in London.

Imperial College refused to confirm that Anderson would be joining its staff, but said that a statement would be released in the "next few days". Anderson has recently faced a series of difficulties at Oxford, including public controversy about his behaviour over a job appointment in the department of zoology (see *Nature* 403, 472 & 404, 214; 2000).

Planets line up for LASCO telescope

Washington The world won't end. But the line-up of five planets on the same side of the Sun, due on 17 May, will at least make a pretty picture. Calling it a 'line-up' is actually a bit of a stretch — Mercury, Venus, Mars, Jupiter and Saturn will appear to be clustered within 19.5 degrees of sky. And the view from Earth will be terrible, as our planet will be on the opposite side of the Sun.

Fortunately, though, the LASCO (large angle and spectrometric coronagraph) instrument on the US/European Solar and Heliospheric Observatory is accustomed to looking in the direction of the Sun, equipped as it is with a coronagraph to block out the



solar disk. LASCO caught three planets in its field of view on 3 May (see above), and will add one more on 15 May (when only Mars will lie outside the frame).

US agency releases its own web language

Paris After HTML and XML, here comes DAML (DARPA Agent Mark Up Language), a web language being developed by the US Defense Advanced Research Projects Agency (DARPA) for advanced searching and distributed computing. The idea behind DAML is to mark up pages in such a way that search engines and other software can read meaning — and not just content — from a page. Software exploring a site written using DAML would be able to map the meaning of the site to any other DAML site.

Getting standards widely adopted on the web is no easy matter (see pages 112–115 and 117–120). But Jim Hendler, a University of Maryland professor on secondment to DARPA, and one of DAML's creators, reckons that its use will be encouraged as DARPA is making its tools available.

Russian ministry urged not to cut funding

Moscow Valery Sobolev, chairman of the trade unions council of the Russian Academy of Sciences, has written to the finance minister's first deputy, Alexei Kudrin, concerning the ministry's plans to reduce science financing in the last quarter of this year. The letter also complains about the draft of a new law prepared by the ministry and recently sent to the State Duma, the lower chamber of the Russian parliament, prohibiting scientific institutions from keeping the money they receive for leasing out their property.

"Taking away the money now collected by scientific organizations for leasing would be a catastrophe for many of them, as would refusing to follow the law on the guaranteed minimum financing of science," says the letter. "Such moves by the finance ministry, if approved by the parliament, would be the

first sign of the coming reduction in science budget financing for 2001, which will only add to the continuing destruction of Russia's intellectual potential."

Complaints force Edinburgh to modify stem-cell patent

Munich The University of Edinburgh last week filed a modification to its controversial patent on a stem-cell technology, which was issued last year by the European Patent Office (EPO) in Munich. The patent came under fire because it includes claims on the alteration of the human germ line. The EPO conceded that the patent was issued by mistake, but said that it does not have the legal right to amend a granted patent by itself (see *Nature* 404, 3; 2000).

According to the reworded patent description — the validity of which will be decided by an EPO appeals board later this year — each of the patent's claims will be restricted to non-human applications. The patent has already been challenged by 7,500 individuals and groups, who have signed a collective objection filed last month by Greenpeace.

Stanford women gain from anonymous gift

San Diego A \$20 million gift to boost the number of women in science and engineering at Stanford University has been made by an anonymous donor. Officials of the university in Palo Alto announced last week that the gift will be used to create endowments to support women fellows in science and engineering. It will also provide a discretionary fund to attract and retain women faculty, and establish an undergraduate engineering diversity fund.

Acknowledging that increasing women faculty "has been slower than we would like in some fields", Pat Jones, a vice provost for faculty development, said: "This gift will certainly enhance our efforts to hire outstanding senior women faculty and... boost the number of promising young women in the pipeline."

Physicist Kelly to head US advocacy organization

Washington The Federation of American Scientists has appointed physicist Henry Kelly as its first new president for 30 years. The federation is an advocacy organization that speaks for many of the United States' leading scientists on nuclear weapons and other science-related issues.

Kelly, who works in the technology division of the White House Office of Science and Technology Policy, will take over the presidency next month from Jeremy Stone, the long-time president of the federation.

Souped-up search engines

For scientists, finding the information they want on the World-Wide Web is a hit-and-miss affair. But, as Declan Butler reports, more sophisticated and specialized search technologies are promising to change all that.

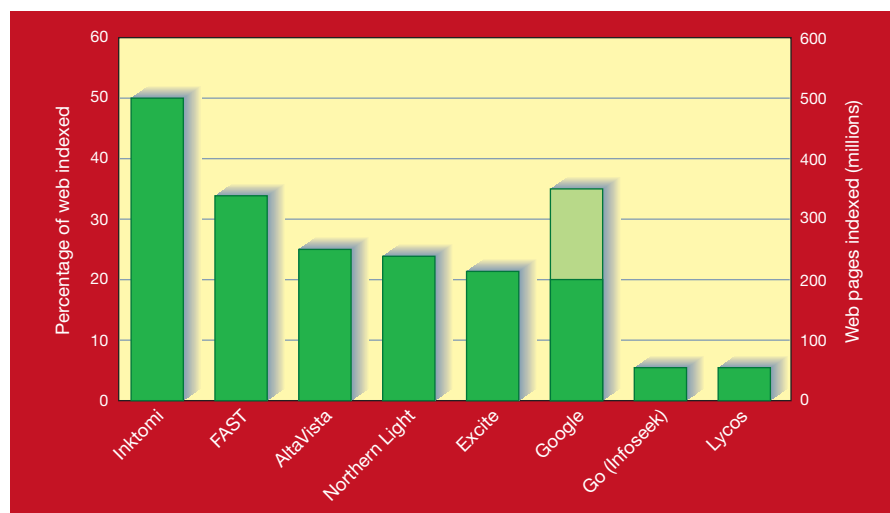
“What is nature?” would seem to be about as vague a question as you can get. But put it to Ask Jeeves, a popular Internet search engine, and its preferred response is clear: “*Nature*, international weekly journal of science”. Search with other leading engines using the single keyword ‘nature’, and *Nature* or its sister journals appear in the top ten returns.

Sadly, many scientific resources on the web are much harder to find. Search engines either miss them entirely, or make you scroll down dozens of pages of hits, your eyes propped open in the hope that pertinent links will leap out from the screen.

Today’s mainstream search engines are simply not run with scientists in mind. Leaving aside their inability to interrogate the vast amount of scientific information held in databases — be they of spectral lines of stars, genome data or events from particle colliders — scientists are poorly served even when they search for text on the web. “There are substantial limitations to search engines and they have bigger implications for scientists than for regular consumers,” observes Steve Lawrence of the NEC Research Institute in Princeton, New Jersey, co-author of a seminal paper on the accessibility of online information (see *Nature* 400, 107; 1999).

Crawling towards the answer

Most search engines rely on ‘crawler’ programs that index a web page, jump to others that this links to, index these, and so on. The starting points bias the end results, and tend to be pages on topics of mass interest — so sport is covered more extensively than quantum computing, for instance. Web pages purposely submitted to search



Falling short: most search engines only cover a fraction of the web. Google's light green bar shows pages it indexes without crawling to. With permission from <http://searchenginewatch.com>.

engines — perhaps by their authors — can also gain prominence. Commercial websites, meanwhile, use a variety of tricks to boost their search rankings, filling their pages with strings of repeated keywords in a colour that makes them invisible to the viewer, or embedding keywords in the HTML (HyperText Markup Language) code that underlies a page.

If search technologies were to stand still, the phenomenal growth of the web would soon render them useless. There are already more than a billion web pages, and even the most wide-reaching search engines cover barely half of these (see figure, above). Within two years, the web may grow to 100 billion pages, and search engines face huge difficulties keeping pace.

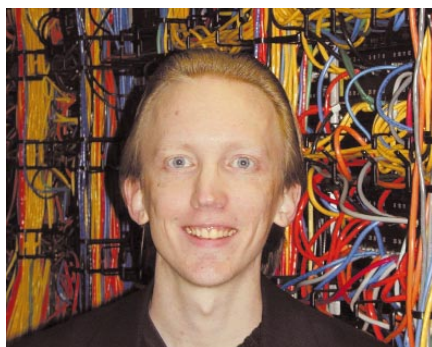
A new study of the structure of the web provides little comfort (see ‘The web is a bow tie’, page 113). This contradicts earlier suggestions that any two pages on the web are connected by a relatively small number of hyperlinks (see *Nature* 401, 131; 1999). The implication is that search engines must crawl from a greater diversity of starting points if they are to have

Rajagopalan: predicts a wave of specialist search engines.

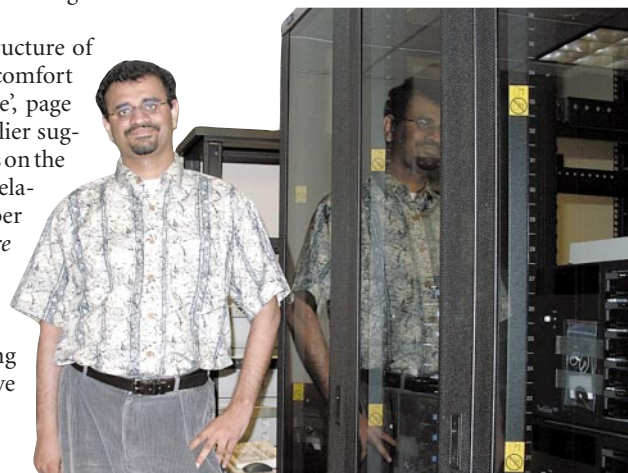
any hope of giving a reasonable breadth of coverage.

Thankfully, new search technologies promise to increase vastly the precision of web searches — and scientists are poised to reap the benefits. The introduction of XML (eXtensible Markup Language), seen as the successor to HTML for coding web pages, should make it possible to restrict a search — for instance to scientific papers, or even to papers that report work using specific biochemical reagents (see ‘The sweet XML of success’, page 114).

Experts predict that within five years, searching the entire web by keywords will be a thing of the past for most researchers. Your



Lawrence: scientific searchers suffer.



personalized search needs may be met by dedicated science search portals. These webs within the web will concentrate many of the online resources you need within an easily navigable environment.

The various online repositories of the scientific literature may by then have adopted common standards to allow seamless searching across them. And the parallel development of 'intelligent' search software could mean that, as you type e-mails or word processing documents, your computer automatically delivers suggestions about relevant web resources, tailored to your particular interests and expertise.

Quality, not quantity

Portals are a hot topic on the web at the moment. The idea is to organize related content so that it can be searched in isolation from the web as a whole. This approach trades off the sheer scale of the available content against quality and ease of navigation. It also allows search engines that are overwhelmed by the public web to perform excellently.

Popular search engines have cottoned on to the trend. The Hotbot engine, for example, lets users search only academic sites with a domain name ending in '.edu'. "Soon you will see a whole slew of search engines specializing in particular sectors," predicts Sridhar Rajagopalan of IBM's Almaden Research Center in San Jose, California.

For the present, however, the biggest innovation in search engine technology takes its inspiration from the citation analyses used on the scientific literature. Conventional search engines use algorithms and simple rules of thumb to rank pages based on the frequency of the keywords specified in a query. But a new breed of engines is also exploiting the structure of the myriad links between web pages. Pages with many links pointing to them — akin to highly cited papers — are considered as 'authorities', and are ranked highest in search returns.

This approach has been pioneered by Sergey Brin and Lawrence Page, two graduate students in computer science at Stanford University in California. In less than a year, their Google search engine has become the most popular on the web, yielding more precise results for most queries than conventional engines — and transforming the lives of its developers. "I haven't finished my PhD," says Brin. "I'm afraid to say I've been too busy with Google."

Google's algorithms rank web pages by analysing their hyperlinks in a series of iterative cycles. "We don't just look at the number of links, but where they come from," explains Brin. "A link from the *Nature* home page will be given more weight than a link from my home page; more things point to *Nature*, therefore it is likely to be more important, and more important things tend to point to



Boy wonders: Brin (right) and Page's Google search engine is the most popular on the web.

Nature, which again suggests that *Nature* is a more important authority."

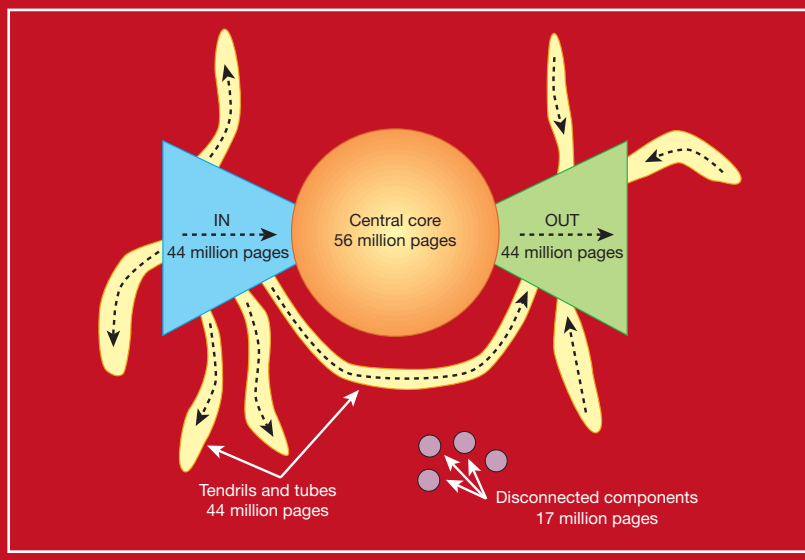
Whereas most search engines only associate the text of a link with the page the link is on, Google also associates it with the page the link points to. This allows it to cover many more pages than it actually crawls, even yielding links to sites that bar search engines' crawler programs.

The web is a bow tie

A study of the web's structure, five times larger than any attempted previously, reveals that it isn't the fully interconnected network that we've been led to believe. The study suggests that the chance of being able to surf between two randomly chosen pages is less than one in four.

Researchers from three Californian groups — at IBM's Almaden Research Center in San Jose, the Altavista search engine in San Mateo and Compaq Systems Research Center in Palo Alto — have analysed 200 million web pages and 1.5 billion hyperlinks. Their results, which will be presented next week at the World Wide Web 9 Conference in Amsterdam, indicate that the web is made up of four distinct components.

A central core contains pages between which users can surf easily. Another large cluster, labelled 'in', contains pages that link to the core but cannot be reached from it. These are often new pages that have not yet been linked to. A separate 'out' cluster consists of pages that can be reached from the core but do not link to it, such as corporate websites containing only internal links. Other groups of pages, called 'tendrils' and 'tubes', connect to either the in or out clusters, or both, but not to the core, whereas some pages are completely unconnected. To illustrate this structure, the researchers picture the web as a plot shaped like a bow tie with finger-like projections.



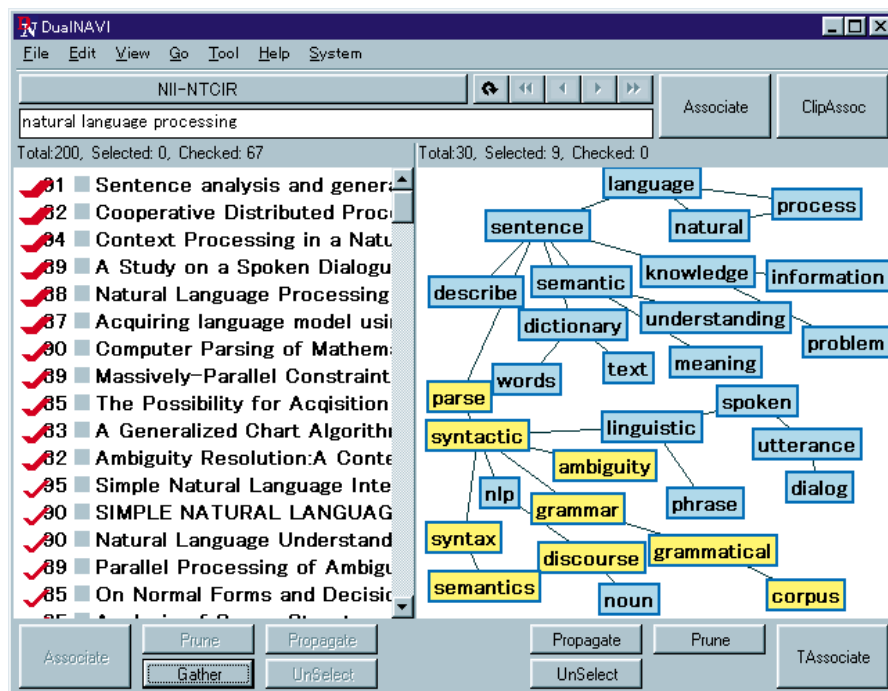
► 'metasearch' engines, such as Sherlock-Hound, which allow users to query multiple search engines simultaneously. The drawback of metasearch engines is that they can give enormous lists of hits. But NEC is working on a next-generation metasearch engine, called Inquirus. Rather than simply spewing out the results from other search engines, it reindexes this initial list to provide a better ranking. Inquirus also checks for broken links, and weeds out duplicate entries.

Searches for researchers

But what about search engines designed specifically for scientists? NEC's prototype, called ResearchIndex, gathers fragmented scientific resources from around the web, and automatically organizes them within a citation index. And unlike most search engines, ResearchIndex retrieves PDF (portable document format) and postscript files, widely used by scientists to format manuscripts. It starts by querying dozens of popular search engines for a series of terms likely to be associated with scientific pages, such as 'PDF', or 'proceedings'. Hundreds of thousands of scientific papers can be located quickly in this way, says Lawrence.

ResearchIndex uses simple rules based on the formatting of a document to extract the title, abstract, author and references of any research paper it finds. It recognizes the various forms of presenting bibliographies, and by comparing these with its database of other articles can conduct automatic citation analyses for all the papers it indexes. This information can also be used to quickly identify articles related to any indexed paper.

The prototype form of ResearchIndex is being applied to the computer sciences. Its archive of papers in this subject alone, at 270,000 articles, is bigger than leading online scientific archives such as the HighWire Press, which has almost 150,000 articles, and the Los Alamos archive of physics preprints, which contains about 130,000 papers. The



Helpful suggestions: DualNAVI generates a graph of potential keywords with which to refine searches.

engine already has an enthusiastic following among computer scientists. Stevan Harnad of the University of Southampton, who has tested the system on his CogPrints archive of preprints in the cognitive sciences, is another convert. "For the literature it covers, it is a gold mine," he says.

NEC is giving the software free to non-commercial users, and Lawrence hopes it will be applied across many disciplines: "Our goal is to not just create another digital library of scientific literature, but to provide algorithms, techniques and software that can be widely used to help improve communication and progress in science."

Concept albums

Other prototype search engines boast features that could make trawling the scientific

literature more efficient. A team at Hitachi's Advanced Research Laboratories, in Hato-yama, Japan, is developing an engine called DualNAVI which could improve the efficiency of searches on collections of scientific literature such as Medline. Hits for a keyword are listed on the left-hand side of the screen. But DualNAVI also generates a set of related keywords by analysing the retrieved documents, and displays these on the right of the screen as a 'topic word graph' (illustrated above). Click any topic, and related documents are highlighted in the left-hand window. This often yields articles that the initial query missed. And in a further twist, groups of documents can be selected, indexed for keywords, and run against other literature databases to find related papers.

Collexis, a Dutch firm based in The

The sweet XML of success

Scientific information would be easier to find on the web if it were clearly marked as such. This is the promise of XML, soon to become the language of choice for web pages.

Current HTML coding tells browser programs little more than how a page should look. XML allows web pages to specify data and what they are, allowing browsers not just to read pages, but to process data referred to in the pages by machine readable tags, or 'metadata'. Using XML, one could, for example, state that a page is a scientific paper, and provide information such as author,

address and keywords. Tags can also represent fields in a database, allowing browsers to interface directly with datasets on the web.

"It would be possible to label a page as being about, say, the Viking missions to Mars, and have specific metadata attached to images that could identify them as being linked to the names of the features they depict," says Robert Miner of Geometry Technologies, a company in St Paul, Minnesota, specializing in web sites for scientific applications.

Some experts are sceptical of any strategy that relies on the entire

web community agreeing formats for tagging information. But in well-organized scientific circles, it should work better. Some disciplines have already drafted their own metadata standards, such as MathML, agreed by the mathematics working group of the World Wide Web Consortium (W3C). At present, mathematical notation is usually represented on web pages using image files, but with MathML it can be described precisely. This would allow researchers to search for pages containing particular symbols. Some software developers are already

developing tools that will generate the metadata automatically.

The humble hypertext link is also set for a facelift. The W3C is developing XLink and XPointer, which will make hyperlinks much more sophisticated. Xlink will let users append their own links to pages on the web, for example, with a single link offering multiple destinations. Unlike today's hyperlinks, XPointer allows links to point to precise paragraphs or sentences, so search engines will be able to return the precise part of a document that seems relevant.

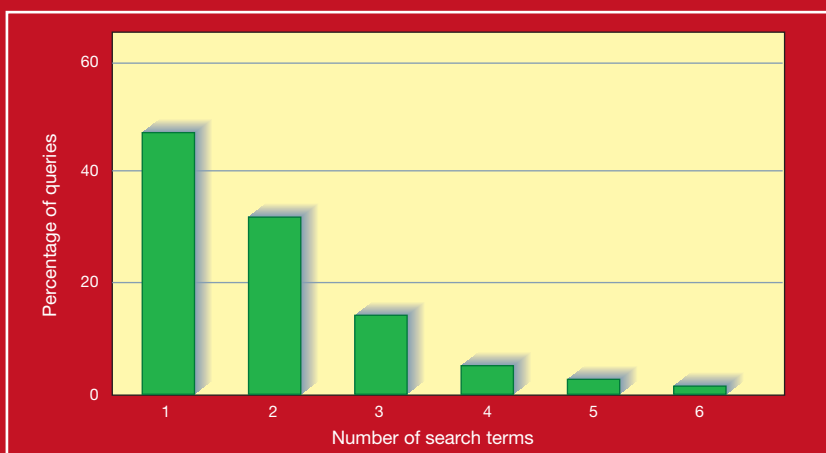
Never trust a human

As search engines become ever more sophisticated, human inadequacies are becoming the biggest limiting factor. A survey by the NEC Research Institute in Princeton, New Jersey, reveals that up to 70 per cent of web users typically type in only one keyword or search term. This is a recipe for obtaining long lists of irrelevant hits. Yet even among the staff of the NEC Research Institute, who should be web-savvy, almost half fail to define their searches more precisely (see figure).

Statistics such as these have convinced

search engine developers that there is little point trying to get users to read the 'search tips' or use the advanced search options provided by most search engines.

To address this problem, more sophisticated engines are now building classifications hidden behind the search interface, which allow them to prompt for more information when presented with ambiguous search terms. Asked to search for 'green', for example, the engine might ask whether the user wants to search for green beans, green parties, and so on.



Hague, has taken this approach further by developing algorithms that can analyse documents not just for keywords but for 'concepts'. Taking its guide from thesauri such as the US National Library of Medicine's Unified Medical Language System, which recognizes more than 600,000 different biomedical concepts, Collexis's engine looks for defined patterns of terms and analyses their contextual relationships.

Users can paste any text, such as a scientific paper, into a search box. This is automatically analysed to identify its major concepts, creating a profile that is then used to search for similar texts. Collexis also allows users to fine-tune their searches by adjusting the weighting given to each individual concept.

Because the Collexis engine's concept matching does not depend on a document's precise grammatical structure, it can generate concepts for texts in many languages, based on a rough machine translation of a document's keywords. Indeed, the software was developed as part of a European Union-funded project to facilitate international cooperation between health researchers. In keeping with these origins, 20 per cent of Collexis's profits will go to promoting collaboration between medical researchers in developed and developing countries.

Autonomy, a San Francisco-based company, is also using concepts. It generates these using neural networks and pattern-

matching technologies developed at the University of Cambridge. Rather than simply responding to specific queries, however, Autonomy is developing programs that automatically analyse any text in an active window on a user's screen to suggest related web resources. The company has released a free, pared-down version of its software, called Kenjin. As you type, Kenjin continuously proposes web resources in a small window at the bottom of the screen. This is still rudimentary, but gives a pointer to the future.

Google's Brin predicts that in five years the search engine as we know it will no longer exist, or be marginal. In its place will come 'intelligent' programs that search by using their experience of the needs and interests of their users. Rajagopalan agrees: "In future there will be automatic feedback loops based on what search results you have selected in the past in relation to this or that query, or how long you stayed at particular web pages."

Learning from previous search sessions, such programs may, in the manner of Autonomy's program, continuously conduct background searches and suggest web resources — with these being tailored to the interests of individual users. Such automated technologies should also help solve one of the biggest limitations to efficient searching: the fact that most people don't frame their queries to maximize the relevance of search returns (see 'Never trust a human', above).

As search technologies improve, journal publishers and the operators of electronic preprint archives are also taking steps that should make it much easier to search for scientific papers on the web. Crossref, a scheme agreed to last November by leading commercial scientific publishers, will link references in the articles they publish to the source papers in their respective publications (see *Nature* 402, 226; 1999). More than three million articles across thousands of journals will become searchable this year, with half a million being added every year thereafter. And in February, the operators of the world's leading archives of electronic preprints, or 'e-prints', agreed standard formats that should allow scientists to search across all of them simultaneously.

This Santa Fe Convention will allow e-prints to be tagged as 'refereed' or 'unrefereed', along with other information such as author and keywords. Using software that will become available this month, any scientist could, in principle, set up an e-print or refereed website, or a site of conference proceedings, in the knowledge that it would be compatible with this global system.

If these efforts succeed in weaving a seamless web from the scientific literature, researchers should find that the hours spent trawling through pages of irrelevant search returns are consigned to history.

Declan Butler is Nature's European correspondent.

An enhanced version of this feature is available online

Search engines

AltaVista ♦ <http://www.altavista.com>

Ask Jeeves ♦ <http://www.askjeeves.com>

Direct Hit ♦ <http://www.directhit.com>

Excite ♦ <http://www.excite.com>

FAST Search ♦ <http://www.alltheweb.com>

Go (Infoseek) ♦ <http://www.go.com>

Lycos ♦ <http://www.lycos.com>

HotBot ♦ <http://www.hotbot.com>

Inktomi ♦ <http://www.inktomi.com/products/portal/search/tryit.html>

Northern Light ♦ <http://www.northernlight.com>

Advanced search engines

Google ♦ <http://www.google.com>

IBM Clever project ♦ <http://www.almaden.ibm.com/cs/k53/clever.html>

ResearchIndex ♦ <http://www.researchindex.com>

Metasearch engine

SherlockHound ♦ <http://www.sherlockhound.com>

Automated search tools

Autonomy ♦ <http://www.autonomy.com>

Kenjin ♦ <http://www.kenjin.com>

Standards

The Santa Fe Convention ♦ <http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html>

W3C XML page ♦ <http://www.w3.org/XML/Activity>

MathML ♦ <http://www.w3.org/Math>

XLink ♦ <http://www.w3.org/TR/xlink>

XPointer ♦ <http://www.w3.org/TR/xptr>

We urgently need more data to improve the lives of laboratory animals

Sir—One of the tasks of the international Organisation for Economic Co-operation and Development (OECD) is to produce standardized guidelines for harmonizing national regulations. The Paris-based OECD's environmental directorate is currently preparing a draft guidance document on the recognition, assessment and use of clinical signs as humane endpoints for experimental animals used in safety evaluation (OECD Series on Testing and Assessment, N.19, www.oecd.org/ehs/test/flags.htm). Experts and interested people are invited to submit comments through national coordinators.

Comments, including ours, have been delivered to OECD in the past few weeks. But we are concerned that some important scientific and ethical issues may not be receiving the attention they deserve. On other occasions we have pointed out some shortcomings in guidelines for the care of laboratory animals. For example, the recent US National Research Council guide is more concerned with dogs than with rodents, possibly because of the influence of veterinarians, who are trained to care for pet animals, in preparing the guidelines¹. Yet rodents are by far the most commonly used experimental species of mammal.

The current OECD document risks overlooking some behavioural features of laboratory animals that could help in improving their general level of welfare. To take one example, behavioural observations aid in judging the vitality of the newborn rodent pup. The degree of spatial dispersion at the nest site provides an assessment that is otherwise difficult in newborns before their eyes are open, but under current guidelines is vital to the declaration of the 'moribund' state which compulsorily precedes euthanasia. In our experience, the ultrasonic calling pattern is also highly informative.

Furthermore, refined ethological techniques (some devised in our laboratory) provide improved, lower-suffering methods for measuring the reaction of rodents toward painful stimuli². These experimental protocols, protecting both data quality and psychophysical welfare, received substantial approval at the third international congress on alternatives in animal use, held in Bologna last summer³.

The care and maintenance of laboratory animals has to take into account the evolutionary history of the species under study. This is particularly true when the subjects are laboratory-maintained male mice or female rats, living all their lives in unnatural

social settings which fairly frequently result in fighting⁴. The need to consider behavioural and species-specific characteristics is also relevant when primates are used as experimental animals. The European Federation of Primatology (<http://www.unipv.it/webbio/efp/efp.htm>), through its primate expert group, is currently preparing a series of recommendations on minimal cage-sizes and specific environmental enrichment techniques, to be proposed to the Council of Europe.

There is still much work to do, and we urgently need more reliable behavioural data. This is particularly true for animals used for neuroscience research, where there is possibly most potential for stress and overt suffering. The ethologist's perspective will also be of great use given the explosion in use of transgenic mouse strains, some of which have unexpectedly been found to be hyperalgesic or to have abnormal levels of fear and inter-male aggressive behaviour.

Enrico Alleva, Augusto Vitale

Laboratory of Physiopathology of Organ and System, Istituto Superiore di Sanità, Viale Regina Elena 299, I-00161 Rome, Italy

1. Alleva, E. & Santucci, D. *Ethology* 103, 1072–1073 (1997).
2. Cirulli, F., De Acetis, L. & Alleva, E. in *Progress in the Reduction, Refinement and Replacement of Animal Experimentation* (eds Balls, M. et al.) (Elsevier, Amsterdam, in the press).
3. Balls, M., van Zeller, A.-M., Halder, M. (eds) *Progress in the Reduction, Refinement and Replacement of Animal Experimentation* (Elsevier, Amsterdam, in the press).
4. Alleva, E., Petrucci, S. & Ricceri, L. in *Behavioural Brain Research in Naturalistic and Semi-Naturalistic Settings* (eds Alleva, E. et al.) 359–374 (Kluwer, Dordrecht, 1995).

Planck family paid a high price for opposing Hitler

Sir—Your News item "German science starts facing up to its historical amnesia"¹ should rivet the attention of the living generation of scientific workers on the tragic descent of broad sections of German science and medicine into the embrace of Hitler and Goebbels.

Under the Third Reich, 45 per cent of German physicians became Nazi party members², and physicists of the rank of Stark and Lenard, together with the world-class mathematician Bieberbach, robustly espoused Nazi doctrine and practice and railed against contamination of the German academy by 'Jewish physics', notably Einstein's special relativity theory³.

Unfortunately, the doctored and undoctored 1934 group photo of Max Planck which accompanies your news item may mislead readers into conceiving Max Planck himself to be a willing fraternizer with German fascism. In fact, Planck and his closest associate, Max von Laue, were among the first to assimilate and respond to Einstein's great 1905 paper. As Einstein's biographer Abraham Pais notes⁴: "The

rapidity with which special relativity became a topic of discussion is largely due to Planck's early interest".

These events in the first decade of the twentieth century were followed by at least three memorable events under the Third Reich. In 1933, during a face-to-face showdown, Planck pleaded in vain with a highly agitated Führer to stop the Nazi purge of the German academy⁵. In 1935 Planck officially defied Hitler by publicly commemorating Fritz Haber at the Kaiser Wilhelm Institute⁶. And, in 1945, the Gestapo executed Planck's last surviving child, Erwin Planck, for complicity in the Stauffenberg conspiracy to assassinate Hitler.

William Steinsmith

239 Castenada Avenue, San Francisco, California 94116, USA

1. Abbott, A. *Nature* 403, 474 (2000).
2. *The Value of the Human Being: Medicine in Germany, 1918–1945* (Ärztammer, Berlin, 1991).
3. *Nature* 133, 290 (1934); 137, 93 & 141 (1936).
4. Pais, A. *Subtle is the Lord: The Science and Life of Albert Einstein 149–150* (Oxford Univ. Press, Oxford, 1982).
5. Heilbron, J. L. *The Dilemmas of an Upright Man: Max Planck as a Spokesman for German Science* (Univ. of California Press, Berkeley, 1986).
6. *Nature* 135, 216 (1935); 158, 170 (1946).

How to make diplomats scientifically literate

Sir—I read with great interest the Opinion article (*Nature* 404, 1; 2000) concerning attempts by the US Secretary of State, Madeleine Albright, to produce a scientifically literate State Department. Such attempts will, of course, fail in the end, since the approach being taken by Albright and State Department officials is at best inconsequential, and at worst, perverse.

Hiring an upper-echelon official merely exacerbates the already deleterious hierarchical structure that exists in the US State Department and does not achieve any objective whatsoever. The clear and obvious way to improve the 'scientific literacy' of State Department officials is, first, to remove the bias against scientists present among many diplomats in the State Department and, second, to hire more (not fewer) scientists as foreign service officers and science attachés at embassies overseas, particularly those scientists who have abundant international affairs experience, knowledge of foreign languages and expertise in global science policy.

Until the prejudiced belief that scientists somehow cannot make competent diplomats is completely removed from the US State Department, actions like Albright's are mere window-dressing.

Peter Cohen

The Herbarium, University of Michigan, Ann Arbor, Michigan 48109, USA

Science's neglected legacy

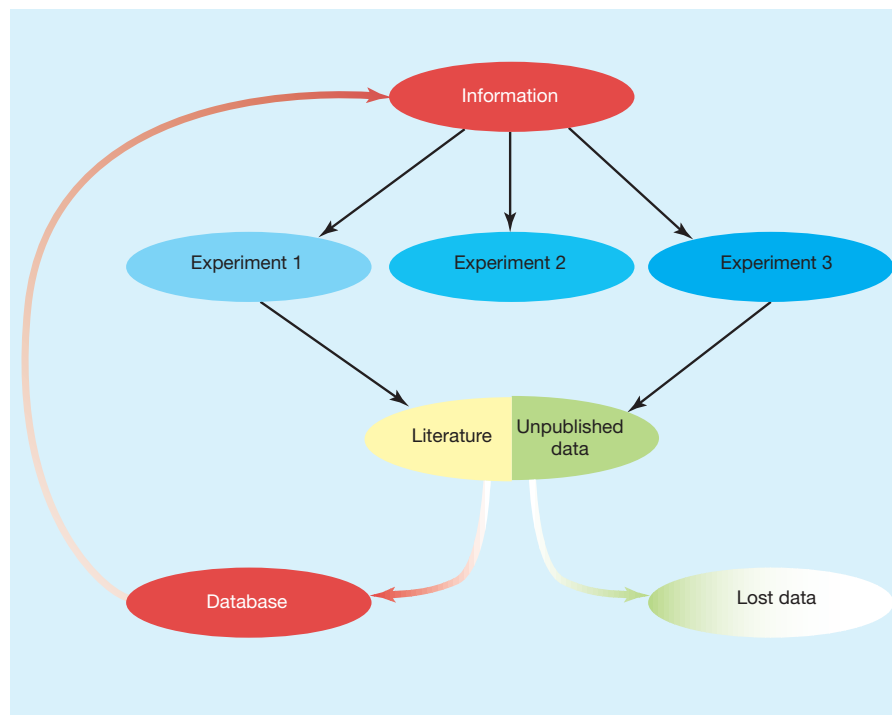
Large, sophisticated databases cannot be left to chance and improvisation.

**Stephen M. Maurer,
Richard B. Firestone
and Charles R. Scriver**

Science is an assault on ignorance. Its legacies are concepts, technologies and databases¹. As with many walks of life, the most glamorous legacies tend to get the most attention and the least are neglected. Databases have never been regarded as glamorous, only useful, and their neglect has not been catastrophic to date because the scientific community is full of individuals and small groups who know how to volunteer, improvise and get things done. Somehow, the most pressing database requirements have always seemed to be met.

But scientific research has outgrown this system. Not long ago, most databases were small (tens of kilobytes) and published as typeset tables or simple online documents. Today, far larger and more complex databases are urgently needed in many fields at a level well beyond the reach of the traditional model of solitary workers or small groups.

For many disciplines, commercial forces are also becoming a problem. This is especially true when the participants in collaborations and government-to-government exchanges are from countries that offer radically different 'property rights' in data (see Box 1). Individuals and small groups are unlikely to bridge such differences on their own. For better or worse, professional societies, government agencies



Databases benefit next-generation experiments by intercepting data that would otherwise be lost.

or other large organizations will have to get involved.

Consequences of failure

If future databases are at risk, what are the consequences of failure? The answer depends on the functions of the databases and how they are expected to perform. Acquiring data is pointless unless society

actually captures what has been learned and uses the information. How well are scientists doing in this regard?

Consider 'big science'. When the Lawrence Berkeley National Laboratory in California shut down its Bevatron accelerator in 1993, most of the facility's unique heavy-ion data had never been published in any form. With better planning, scientists could

Box 1: Commercial turmoil

It is not hard to imagine how the rise of commercial forces in certain fields — notably biotechnology — could interfere with databases' ability to collect and publish data. But will it? So far, the main problem posed by commercialization is not so much greed as uncertainty.

Academic nervousness. Very few academic scientists know what their data would be worth on the open market. Despite this, scientists who operate databases often hear colleagues complain that data-sharing agreements could give valuable information away 'by accident'. It is difficult to say how often such statements translate into actual behaviour. Existing empirical evidence suggests that the effect is still fairly modest⁹.

New ways of doing business. The scientific-database business is so unfamiliar that entrepreneurs are still experimenting with new kinds of contracts. For example, Swiss-Prot often claims 'pass-through' rights in its data even after the information has been re-worked and merged with other material to create new databases (see http://www.expasy.ch/announce/sp_98sum.html for Swiss-Prot's licensing policy). Similarly, Celera has announced that users will not

be allowed to redistribute its data¹⁰. Such practices could limit scientists' ability to create new databases that build on and extend existing resources.

Legal uncertainty. In 1996, the European Union created strong property rights in data; similar, but weaker, legislation is pending in the United States.

Compromise seems inevitable, but may not happen for years. In the meantime, some US scientists have reported that their European collaborators are reluctant to share data. Government-to-government weather-data exchanges have also been affected.

It has been claimed that an analogous mix of inexperience and uncertainty has already interfered with biotechnology's ability to license and exploit patented inventions¹¹. Commercial forces may be acting as a similar brake on the creation of databases. The good news is that such problems should disappear over time. According to economists, excessive secrecy is irrational because owners can always make more money by entering into profit-sharing agreements and then pooling their data¹². In the long run, commercial forces should actually reinforce the traditional ethos of sharing information.

still be using this information to supply critical measurements in hot areas such as solar neutrinos, nucleosynthesis and cosmic rays. Instead, they will probably have to wait decades before these data are remeasured.

Now, history looks set to repeat itself. The US Department of Energy is currently developing advanced information systems for two new \$600 million accelerators — the Relativistic Heavy-Ion Collider at Brookhaven National Laboratory, New York, and the Thomas Jefferson National Accelerator Facility, Virginia. Unfortunately, neither system is designed to archive and disseminate data over the long run, and the energy

department, although aware of the problem, does not have the money to address it.

In the life sciences, millions of observations about location, interpersonal variation and function within the human genome are produced but not published. In principle, these data contain important clues for evolutionary biologists, biochemists, neuroscientists and health-care personnel, to name a few. Highly automated central depositories could make these data available to everyone.

Serving new users

Society cannot get full value for its investment in science unless anyone desiring

existing data actually gets them. So far, success has been uneven. In many fields, the prevailing attitude seems to be that academic databases are 'good enough' for everyone. That misses the point. When non-specialists can't use a particular database, the data might just as well be lost.

Some fields, such as chemistry (see Table 1), have long traditions of re-tailoring data so that they can be used elsewhere. Unfortunately, the experience in physics is more typical. Extensive government funding has allowed nuclear physicists to produce arguably the best scientific databases anywhere. Despite these advantages, most

Table 1 **Mainstays of science**

Database	Data stored	Maintenance
Review of Particle Physics/ Particle Data Group (1957–present)	Mass, spin and other properties for several hundred elementary particles, plus limits on many hypothetical particles.	Government-supported online database; a printed version is regularly published in various journals. http://pdg.lbl.gov
Evaluated Nuclear Structure Data File (ENSDF) (1946–present)	Fundamental nuclear structure and decay information for 3,100 nuclei and isomers.	Database maintained by Brookhaven National Laboratory's National Nuclear Data Center. Published online and as printed Nuclear Data Sheets. Government-supported with supplementary funding by Academic Press. http://www.nndc.bnl.gov/nndc/ensdf/
Table of Isotopes (1940–present)	Advanced, user-friendly listing of ENSDF, including flexible research tools	Government-supported online database maintained by Lawrence Berkeley Laboratory. http://isotopes.lbl.gov/isotopes/toi.html A version with extended content is published commercially by John Wiley.
Swiss-Prot (1986–present)	Protein sequence data, functions, domain structures, variants and related information.	Publicly supported database until 1996. Currently supported by commercial licensing fees plus additional funds from the Swiss government, the European Bioinformatics Institute and other public bodies. http://www.expasy.ch/sprot/sprot-top.html
Chemical Abstracts (1907–present)	Experimental data describing more than 22 million substances.	Non-profit entity supported by the sale of printed, electronic and online databases to academic researchers, industrial scientists, engineers and patent lawyers. http://info.cas.org/casdb.html
GenBank (1982–present)	Community-wide depository listing approximately 4.6 million sequences (more than 3 billion base pairs) found in the human genome.	Government-supported database available free over the Internet. http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html
Nuclear Astrophysics Data Center (proposed)	Regularly updated, electronically searchable nuclear-reaction-rate database tailored to meet the needs of finite-element modellers.	Government-supported extension of existing databases and unpublished data. A limited version of the proposed website can be found at http://ie.lbl.gov/astro.html
Human Mutations Data	Advanced architecture database unifying and extending more than 150 separate websites.	Community-wide database to be launched as a private/public consortium.
DbSNP (1998–present)	Central depository containing more than 2.6 million single-nucleotide polymorphisms (SNPs). Advanced software lets scientists deposit thousands of SNPs at a time.	Government-supported database maintained by the US National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/SNP/
PathoSeq	Gene-sequence data focusing on disease-causing bacteria and fungi. Advanced relational database supports powerful search tools.	Proprietary database sold to industrial subscribers; primarily used in pharmaceutical research. http://www.incyte.com/products/pathoseq/pathoseq.html
SenseLab (1993–present)	Advanced data warehouse combining four previously independent neuroscience databases into a uniform, seamlessly searchable resource for neuron properties and models.	Government-supported academic research tool and technology development project. http://ycmi.med.yale.edu/senselab/
Distributed Active Archive Centers (DAACs)	Archived data sets obtained by Earth-observing satellites, aircraft, field campaigns and other experiments.	NASA-led federation of 10 government data centres, each of which publishes data on a particular topic (for example, "snow and ice", "land processes") relevant to global-change research. Data sets are available online and/or on CD-ROM.
Internets (1996–present)	Search engine providing links to hundreds of scientific databases.	Commercial website. http://www.internets.com/sciencedata.htm
Online Mendelian Inheritance in Man (OMIM)	Annotated catalogue of (largely) Mendelian diseases and polymorphisms; mitochondrial disorders are also recorded.	Community-driven database curated and maintained at the US National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/Omim

Large, electronically searchable databases have become essential to progress in many fields. Increasingly, scientists are experimenting with non-profit and commercial models to supplement traditional funding sources.

nuclear data used in medicine and industry are badly outdated. At least one non-destructive chemical-analysis technology, that of neutron-activation analysis, has already been delayed 30 years by this problem.

If anything, the problem is worse in biology. During the 1990s, private-sector bioinformatics companies spent much time turning academic databases into commercial research tools. Despite this investment, many hot research areas are still poorly served. For example, manufacturers of silicon microarray devices urgently need a community-wide database of reported observations to understand and improve their technology. Similarly, researchers studying single-nucleotide polymorphisms (SNPs) would like to compare their data against a comprehensive database describing all known human mutations. But the required databases simply don't exist^{2,3}.

Discovery does not end with publication, because the literature often contains overlapping and contradictory results. Far from being a nuisance, this discrepant information can be exploited to identify errors and recommend best values.

Modern technology has revolutionized this process in two crucial respects. First, on-line databases are never fixed in stone. When questionable new data appear in the *Table of Isotopes*, for instance, affected authors are quick to send e-mails and unpublished information. This practice provides an important safeguard against editorial errors. The same is true for biological databases.

Second, computers have systematized the traditional editor's trick of doing 'sanity check' calculations to see if a particular result looks reasonable. For example, *Table of Isotopes* evaluators routinely combine existing results to calculate additional variables. If the new variable conflicts with known data, the result provides a powerful tool for detecting errors in one (or more) of the input measurements. Otherwise, the new quantity can be added to the literature as a result in its own right. Over the past five years, *Table of Isotopes* has discovered thousands of discrepancies and many new data in this way. Similar opportunities exist in biology, where creators of many databases (for example, for genetic mutations) already have a wide variety of consistency-checking software products to choose from.

An obvious extension of this strategy is to make observations 'in silico'. Several US and European initiatives are working on enormous (up to 40 terabyte) 'virtual sky' archives for astronomers to explore. Early versions of these projects have already found a rich harvest of quasars and other objects^{4,5}. Beyond virtual observations lie virtual experiments. Here, too, accurate data are essential. For example, supercomputer

Discrepant information can be exploited to identify errors and recommend best values.

simulations of controlled fusion, stellar evolution and supernovae are no better than the physics that goes into them. Unfortunately, there is still no up-to-date, curated database that delivers this information in the formats used by modellers.

Purists may object that inferences from the literature and supercomputer simulations are never as accurate as a good experiment, but that is rarely the choice. Many

experiments are unaffordable; others are beyond existing technology. Under these circumstances, using large databases as if they were virtual particle accelerators (or observatories, or gene-sequencing machines) may be the only available option.

Data mining

Automated surveys for faint signals have always depended on advanced databases. Until fairly recently, most of these 'rare event' searches were limited to high-energy physics experiments. By the late 1990s, however, similar searches had become 'grand challenges' in such diverse fields as astronomy, climatology, meteorology, Earth imaging and biology.

Thirty years ago, constructing a computer-searchable database big enough to hold 1.5 million events (several hundred megabytes of data) per year was enough to win Luis Alvarez of the Lawrence Berkeley National

Box 2: E-solutions?

Modern scientific databases are labour-intensive. Despite aggressive computerization, database creators still do much of their work by hand, one entry at a time. Because automation has been under way for nearly half a century, conventional techniques are approaching their limits. This leaves new technology, especially on the Internet, as the most likely way forward. Could future web tools strengthen — or even replace — current methods for producing databases?

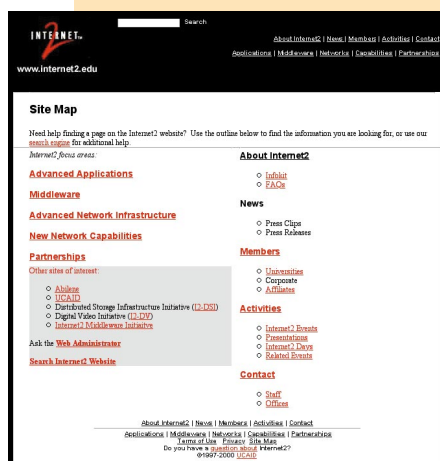
New web tools will make database creation easier in at least three ways. First, scientists need web crawlers to find online data. However, today's crawlers are still far less complete than old-style paper bibliographies. Web developers are working hard to close this gap. Second, some large databases (notably GenBank) are starting to outgrow the web's ability to store and transmit information. CERN's GIOD and The US Grand Challenge collaborations will ease this threat by developing technologies that distribute data across multiple servers (ref. 4 and <http://chep2000.pd.infn.it/paper/pap-c292.pdf>).

Finally, 'middleware' tools are making it easier for cooperating researchers to link the data on their servers into a single, overarching network. One of the best current examples is the US Geological Survey's National Spatial Data Infrastructure Clearing House, which allows users to find and retrieve climate data stored on more than 200 participating servers (J. Restivo, personal communication). Better middleware tools are a top priority for the US Internet 2 and CERN's Grid initiatives (see <http://www.internet2.edu/middleware/overview/areas-of-activity.html>).

Good as they are, none of these tools replaces human editors when it comes to combining and evaluating independent databases. How likely is this to happen? In one promising strategy, human editors write detailed instructions for translating each database's unique computing

conventions, data fields and scientific nomenclatures into a standard format. Computers then execute the instructions to build a 'data warehouse'. Unfortunately, current warehouses are rarely able to unify more than a dozen databases. After that, the amount of human intervention required to create and update the instructions becomes unworkable (T. Slezak, personal communication). In principle, advanced web crawlers could break this impasse by learning to translate data on their own. For now, though, crawlers still have trouble retrieving even simple objects such as telephone listings. Most Internet researchers agree that human editors are not likely to be replaced soon.

New web tools could help put large, sophisticated databases back within the reach of individuals and small groups. At the same time, technology is not enough. For the foreseeable future, even the most powerful web tools will depend on humans to organize networks, adopt standard nomenclatures and combine independent data sets. New tools are only part of the solution. The biggest challenges are social and economic.



Laboratory the Nobel prize for physics. Today, experiments such as Lawrence Berkeley's STAR detector can record 54 terabytes ($\times 10^6$ megabytes) per year. Experiments at the Large Hadron Collider at the European Organization for Nuclear Research (CERN) are expected to reach 100 petabytes ($\times 10^9$ megabytes) by roughly 2010 (ref. 4).

Meanwhile, biologists are already planning for the day when silicon microarray detectors can simultaneously measure the 'expression level' for each of the body's estimated 100,000 genes. Exploiting these data will mean finding the handfuls-to-hundreds of genes responsible for particular biological processes and diseases. Scientists are actively discussing the statistics and computing methods needed to separate these tiny, complicated signals from the background noise of several thousand genes performing unrelated functions. Whichever strategy is chosen, however, one fact is already obvious: biologists will have to combine observations from hundreds and perhaps thousands of different researchers into a unified, seamlessly searchable database. This challenge is far from trivial in a world where bioinformatics specialists still have trouble combining more than a dozen sites into a unified 'data warehouse'. Commercial pressures to hoard data could complicate the task even further (see Box 1).

Taming the web

Ten years ago, many scientists thought that the World-Wide Web would encourage individuals and small groups to build more databases. In a way, this prophecy came true, and some scientists have even built substantial research agendas around data spotted on the web. Unfortunately, the web's strengths are also its weaknesses. Most researchers need to combine and unify data from as many sources as possible. By contrast, the web usually fragments information by encouraging each scientist to create his or her own home page. In genetics, for example, more than 150 sites are currently devoted to mutations in particular human genes (see <http://ariel.ucs.unimelb.edu.au:80/~cotton/mdi.htm>). In many fields, non-standard computing conventions, nomenclatures and quality-control practices create additional difficulties.

New web tools may ease, but will not eliminate, these problems (see Box 2). For example, many scientists want to combine

Society can't get full value for its investment in science unless anyone desiring existing data actually gets them.

Working Group Draft 2
December 23rd, 1999

HUGO/MDI COMMUNITY-WIDE DATABASE: PROJECT SPECIFICATIONS

A. INTRODUCTION

Mutation databases are becoming increasingly consulted by clinicians and those in the diagnostic arena as evidenced by increasing use of locus-specific databases (LSDBs) and HGMD, a listing of published mutations. There are approximately an equal number of unpublished mutations as published and it is becoming more difficult to publish mutations. The HUGO Mutation Database Initiative (MDI) was formed in 1994 to ensure collection of all mutations in an accurate and reliable manner and their subsequent distribution through the Internet. This task becomes more important as biotechnology later uncovers more mutations and needs the information for progress.

Following extensive discussion, the October 19, 1999 MDI meeting passed a resolution committing the Community to study and, if feasible, implement a new institution designed to (i) strengthen existing LSDBs, and (ii) put integrated, next generation data-sets and software tools into the hands of researchers. The proposed facility that might be dispersed made of many components, would support itself by fundamentally new methods that may include sales to the commercial sector. The proposed facility would provide basic data and research tools to Community members (including LSDBs) without charge.

The first step is to create detailed specifications describing the proposed facility. To this end, a working group including leading bioinformatic experts, LSDB operators, central database operators, and an industry representative was formed. This document reflects their work.

MDI's consultants will use the final version of this document to prepare a detailed business plan. The plan will be presented and put to a vote at MDI's April 9, 2000 meeting in Vancouver. In view of this deadline, it is important that we receive your comments on or before January 31, 2000. Stakeholders and other relevant parties will then be canvassed for expression of interest in forming components of the facility.

B. GENERAL

Members of the Mutation Database Initiative want to build a single depository for their data.

existing climate, biodiversity and environmental databases into large networks. Unfortunately, advanced search engines and data mining are no better than the information they operate on — and existing databases are riddled with errors. Human editors will have to clear the way before these tools can succeed⁶.

We believe that the web's problems are more social than technological. Over the past six years, the Human Genome Organization's Mutation Database Initiative (MDI) has helped to build a broad community consensus in favour of standardized nomenclatures and — to a lesser extent — computing conventions and systems for cross-referencing human mutation data with other types of databases and/or different species^{7,8}. Many MDI members believe that the next step is to build a single, community-wide depository for its members' data. Private companies — which have their own reasons for wanting larger, more unified mutation databases — will be asked to join the project^{2,7,8} (the depository proposal is described at <http://ariel.ucs.unimelb.edu.au/~cotton/specs.htm>).

Changing the system

Most of the needs discussed here are not subtle. They require money. Those involved need to plan for databases in the same way that we think about new buildings and hardware. Several US federal agencies, including the space agency NASA and parts of the Environmental Protection Agency, have taken an important first step by requiring would-be grant recipients to explain how they plan to store and disseminate their data. In biology, similar foresight has been displayed by the launching of several high-visibility efforts to recruit and train more bioinformatics experts. Such initiatives are encouraging, but they are only a start.

Meanwhile, grants from private and public agencies are approaching their limits.

Sales to the commercial sector offer one way out, although this raises the thorny issue of public access. One of the most appealing strategies is for scientists to keep their general research databases in the public domain while selling customized spin-offs to the private sector. The main advantages of such a strategy are that it would avoid commercializing and/or monopolizing basic data, reward science for at least some of its contributions to the economy, and encourage academic scientists to find users who are currently underserved².

Stephen M. Maurer is an intellectual-property attorney at 2632 Hilgard Street, Berkeley, California 94709, USA, and a consultant to the MDI (e-mail: maurer@econ.berkeley.edu). Richard B. Firestone is a staff scientist at the Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA, and author of Table of Isotopes. Charles R. Scriver is Alva Professor of Human Genetics and Professor of Pediatrics and Biology at McGill University-Montreal Children's Hospital Research Institute, Montreal, Quebec H3H 1P3, Canada. He is a co-director of the MDI and curator of the PAHdb knowledge base.

- Ridley, M. *The Economist* **318**, 7644 (16 Feb. 1991).
- Maurer, S. *Hum. Mutat.* **15**, 1 (2000).
- Brazma, A. et al. *Nature* **403**, 699–700 (2000).
- Szalay, A. et al. *Accessing Large Distributed Archives in Astronomy and Particle Physics* http://pcunn.citethp.caltech.edu/aldap/kdi_proposal.htm (1999).
- Murtagh, F. *The Virtual Observatory: Methodologies for Data Handling* <http://newb6.u-strasbg.fr/~ccma/vol/virtual-observatory.html>
- Winker, K. *Nature* **401**, 524 (1999).
- Cotton, R. G. H., McKusick, V. M. & Scriver, C. R. *Science* **279**, 10 (1998).
- Cotton, R. G. H. & Kazanian, H. H. Jr (eds) *Hum. Mutat.* **15**, no. 1, spec. issue (2000).
- Campbell, E. et al. *Res. Policy* **29**, 303–312 (2000).
- Butler, D. & Smaglik, P. *Nature* **403**, 231 (2000).
- Heller, M. A. & Eisenberg, R. S. *Science* **280**, 698 (1998).
- Scotchmer, S. J. *Econ. Perspectives* **5**, 29 (1991).

Acknowledgements. We thank Suzanne Scotchmer, University of California at Berkeley, Tom Slezak of the Joint Genome Institute/Lawrence Livermore National Laboratory and Richard Cotton, Mutation Research Centre, St Vincent's Hospital, Melbourne, Australia, for suggestions and support.

Battle of the bones

A feud between two palaeontologists sheds light on late-Victorian journalism.

The Bonehunters' Revenge: Dinosaurs, Greed, and the Greatest Scientific Feud of the Gilded Age

by David Rains Wallace
Houghton Mifflin Company: 1999. 347 pp.
\$25

Kevin Padian

From about 1868 until their deaths in the 1890s, two American vertebrate palaeontologists, O. C. Marsh and E. D. Cope, carried on a personal and professional feud that has dogged the field ever since. Even today it seems that whenever a controversy about the interpretation of ancient bones arises, journalists gravitate towards the personal side, and eventually raise the spectres of Cope and Marsh. The exploration of those animosities and the journalistic ethics that fuel them set David Rains Wallace's new book apart from much of the rest of the literature on this well-trodden subject.

The story takes place during the "Gilded Age" — as the wealth-obsessed, politically corrupt period between the American Civil War and the First World War is ironically known — which creates a vivid backdrop to the antics of these prominent scientists. The basic facts of the feud are generally understood, although previous biographers have emphasized or interpreted them differently. Cope was the scion of a wealthy Philadelphia Quaker family; Marsh was the poor relative of an expatriate American millionaire, George Peabody. Cope was a "master naturalist", as his acolyte and biographer, the palaeontologist H. F. Osborn, pronounced him. But Cope was also vain, delusional, belligerent and incredibly naive about the workings of society. Marsh is frequently characterized as methodical, slow and simultaneously plodding and plotting in his efforts to build his influence in science and politics.

Both men used inherited wealth to build their careers, but their diametrically opposite temperaments, vast ambition and intense competition led the once-fast friends into a bitter rivalry that eventually exploded on the pages of newspapers and the halls of Washington — to the eventual detriment of the entire field of geology and palaeontology. Cope made wild accusations that Marsh had misappropriated funds and influence in the US Geological Survey. Although no evidence was ever found for misdeeds by either Marsh or the survey director J. W. Powell, the scandal caused the sharpest decline in the survey's congressional funding until Newt Gingrich's Republican Revolution of the 1990s.

In Wallace's book, the now-familiar

squabbles on the outcrops of Wyoming and New Mexico, the accusations of plagiarism, fraud and theft, and the ultimate exposure of the feud to the pages of the *New York Herald* are ably reviewed. But Wallace goes further to provide a new and useful perspective as a journalist and writer.

Most studies of the affair have focused on the conspiracy between Cope and William Hosea Ballou, a freelance journalist, to 'expose' Marsh's offences in the pages of the *New York Herald*, making Ballou an all-too-willing conduit for Cope's venom. Wallace digs deeper into the journalistic politics and ethics of the time and provides convincing circumstantial evidence that the *Herald's* monomaniacal publisher, the sociopathic socialite James Gordon Bennett Jr, was behind every word and attitude ever printed in his paper.

Bennett's tyrannical behaviour in the newsroom was surpassed by his social out-

rages. As Wallace puts it, "On New Year's Day of 1877, he drunkenly urinated in the fireplace (or grand piano) of his fiancée's parlor during a gala party. For this, he was horse-whipped by the fiancée's brother and permanently banished from New York society." Yet Bennett had a sharp sense for what entertained the public. He was responsible for sending Henry M. Stanley to Africa to locate Dr David Livingstone, and was keenly aware of how reports of exploration could captivate an audience in the age of Manifest Destiny — the belief in the inevitable westward expansion of US boundaries to the Pacific. He knew how to fan a smouldering scandal, and he loved to inflate and deflate authorities — often in a single article. Wallace regards him as perhaps the most underestimated man of the late-nineteenth-century United States, a driven, tightly wound alcoholic in whom control manifested itself most strongly in his drive to dominate his profession. This

The bones of modern America

The painter Georgia O'Keefe spent 25 years making animal bones a major theme of her work. She would collect bone specimens during her walks in the New Mexico deserts, and claimed that her skull paintings expressed what she felt about the desert. Her *Horse's Skull with Pink Rose*, shown here, was also called "Life and Death". O'Keefe is one of many artists, among them Alfred Stieglitz, Charles Sheeler and Joseph Stella, whose contributions to the creation of an art that was both American and modern are considered in



The Great American Thing: Modern Art and National Identity, 1915–1935 by Wanda M. Corn (University of California Press, \$50, £30).

monster, Wallace claims, was directly responsible for the reputation for infighting that hounds science to this day.

Wallace's research brings new historical insight to old tales, but the moral doesn't end there. A journalist who understands the mixing of business and personality as well as he does can't resist bringing the story to the present, and in two closing vignettes Wallace writes postscripts to the late Victorian saga. He first recounts the confusing and generally misunderstood fate of Cope's own bones after his death, attempting to sift through the urban legends of palaeontology to trace their true whereabouts. Cope apparently willed his brain and skeleton to science; the skeleton, after passing from the Anthropometric Society to the Wistar Institute in his native Philadelphia, eventually wound up in Philadelphia's Academy of Natural Sciences. The anthropologist-poet Loren Eiseley kept it in his office when he realized what it was, and at some point Cope's lower jaw was lost and doubts were raised about the association of the skull with the body. Mixed in with this tragicomic tale of curation are stories about Cope's alleged desire to become the type specimen of *Homo sapiens*, the vain attempt to designate him as such, and the misadventures of his remains as they travelled the world in the 1990s in the company of a photographer who was writing a book about contemporary dinosaur-hunters.

Less successful is Wallace's attempt in the closing chapter to draw parallels between the Cope-Marsh feud and the disagreements about sociobiology and other topics that have arisen between today's scientists such as E. O. Wilson and S. J. Gould. Cope and Marsh, human examples of the competitive-exclusion principle in ecology, were after fossils and power. Wilson and Gould, like many scientists, have philosophical differences but are in non-competing fields, and their disagreement is far less personal. Their common causes and values range from the promotion of conservation biology to the opposition to 'creation science'.

Bennett's Gilded Age audience would have yawned. But today's readers are likely to be absorbed by Wallace's well-written historical account which, perhaps for the first time, shows that, although scientists of the Cope-Marsh period have been portrayed as warped and deceitful, the journalists of the time were at least as bad. There seems to have been no distortion of fact, no manufactured conversation, no false attribution of quotations to which Bennett and his crowd would not have stooped. The book must be successful if it can give a contemporary scientist such as myself a measure of comparative gratitude for the journalistic ethics of our own time. ■

Kevin Padian is in the Department of Integrative Biology, University of California, Berkeley, California 94720-3140, USA.

An ignoble lineage

Cancer: The Evolutionary Legacy

by Mel Greaves

Oxford University Press: 2000. 276 pp.

£19.99, \$27.50

Leo Kinlen

Mel Greaves gives us a laboratory scientist's view of cancer in its historical, epidemiological and experimental aspects — but, unusually, from an evolutionary perspective enlightened by molecular biology. The thesis that cancer is a legacy of evolution must be a truism, since, if we have come to be what we are through evolution, how could cancer not be something 'left' to us by this process? Greaves, however, goes further than this, and raises high hopes at the outset with large (and sometimes enigmatic) claims that the fog is lifting as an evolutionary perspective elucidates many of cancer's puzzles.

The potential for clonal expansion, we are told, exists because certain cells have to remain only partially differentiated in order to replenish other cells and tissues. Inherent in this potential, however, is the possibility of malignancy, the territorial expansion of a mutant clone, and this is described in graphic terms. Invasive cancer occurs when a dominant clone of this type emerges, mutinous in character, stone deaf to intercellular social dialogue, immortal and itinerant. And because it is a disease mainly of post-reproductive life, it is not subject to any important natural selective forces.

Tobacco as a cause of cancer, as well as the history of smoking, is covered. Another chapter describes the nineteenth-century epidemic of scrotal cancer in chimney-sweeps who often began work as boys. A chapter on women's cancers is much longer than that on

men's — not from any bias, but because of the differences in our states of knowledge. Breast cancer is seen as a major "own goal", the hormonal penalty of uncoupling sex and reproduction. Several aspects of cancer that are not always appreciated are well brought out, including its multi-stage nature, the importance of chance at each stage in its causation, and the fact that it is not one disease but many.

Greaves has summarized, pithily and readably, a great deal of work, much of it distant from his own field. Understandably, therefore, there are inaccuracies: the average age at breast cancer diagnosis is not 50, but over 60. Even as a crude rate, cancer is not ten times more common than it was a century ago. Indeed, apart from in the elderly, in whom death certification of cancer has improved, age-specific mortality rates from cancer as a whole are now lower than they were in 1911 (the earliest available), and, but for lung cancer, would be much lower. Some of the judgements are surprising; for example, that the evidence for dietary causes of breast cancer is "persuasive" (many would restrict this assessment to obesity in relation to post-menopausal disease); that environmental chemicals must make a "significant contribution to the toll of breast cancer" (the evidence for any effect is at best controversial); that sex is a likely cause of prostate cancer (most epidemiological studies are negative).

"Epidemiological sleuthing can help identify exposures," the author states, but "understanding the causal mechanisms involved and the real why and how it happens is only possible with some biological archaeology" in the laboratory. In fact, much of our knowledge of the causes of human cancer has come from epidemiology, and it is worth remembering also that prevention is sometimes possible without detailed knowledge of mechanisms — for example, by removing asbestos or aniline



Cancerous ills: prevention is sometimes possible without deep knowledge of the mechanisms involved.

dyes from the workplace, or by not smoking.

Greaves has strong views about "lifestyle factors" in relation to cancer causation, partly because it has overtones of "blaming the victim". (Although he does not mention this issue in his chapter on smoking.) Instead, he considers that many cancers are "products of social engineering over which any one of us exercises limited deliberate or informed choice in the normal course of events". This he embraces under the term "the social ratchet", which I found puzzling and confusing. I cannot understand the essential difference between these positions, as all epidemiologists would agree that social factors may make choice more difficult.

An evolutionary perspective, it might be thought, could help forecast something of the future, and a final chapter is devoted to this. The author is no pessimist and foresees new treatments, including 'designer' drugs that will more accurately attack the cancer clone. Surprisingly, the prospect of identifying more infective causes of cancer is not mentioned, although viral causes of two of the world's most common malignancies (cervical and primary liver cancers) were identified only recently.

The book is well written, full of vivid metaphors, succinct descriptions and arresting headings. I found it persuasive at the cellular level, and enjoyed the accounts of specific causes of cancer. However, the evolutionary perspective did not seem to me to provide a useful framework for investigating the causes of cancer or for its prevention, nor were any specific novel measures proposed. Perhaps this perspective is too long-sighted and the concept of "the social ratchet" too vague to be helpful here. ■

Leo Kinlen is in the CRC Cancer Epidemiology Unit, University of Oxford, Radcliffe Infirmary, Oxford OX2 6HE, UK.

A sermon on the mounts

Sedimentology and Sedimentary Basins: From Turbulence to Tectonics

by Mike Leeder

Blackwell Science: 1999. 592 pp. £35, \$69.50

Chris Paola

Among the accumulated scraps of Earth's tectonic system is an immense library of sedimentary rocks, stacks of former planetary surfaces. The oldest entry dates back some 3.8 billion years. The scriptoria for these ancient books are mostly low places: alluvial plains, lakes, deserts and ocean depths. The archive is run by tectonics, which, by slowly driving down the basement, allows the pages to accumulate. For the most part, the archive

is read as it is destroyed, when tectonics raises the ancient books from their depths and opens them to us via erosion.

If the keeper of the vaults is tectonics, the hand that writes most is fluid flow. The ambitious goal of Mike Leeder's engagingly written and well-produced new book is to take us from the details of turbulent flow all the way up to tectonics, and in the process show us how the sedimentary record is created. He is thorough without choking on the details, and comprehensive without simply cataloguing the outcomes of as many studies as he could find. Beyond its remarkable scholarly range and depth, this sequel to Leeder's popular text *Sedimentology* (Chapman & Hall, 1982) is also a lively and informal read.

Leeder makes no secret of having been seduced by the enigmatic beauty of the sedimentary record. His passion shows in the writing, and in the prose and poetry that begin each chapter. (He has been criticized for this. Is sedimentary geology some kind of haven for grinchers? Geology is one of the most romantic of scientific disciplines; the literature that graces Leeder's book illustrates this, and gives the book a soul.)

Sedimentary geology is a descriptive enterprise. The sedimentary record ranges from exquisite detail to yawning gaps: one can marvel at the delicately preserved traces of a Precambrian rain shower, then lose a hundred million years of record just by walking a few steps up the section. This makes it hard to piece together the history of the Earth's surface even at a qualitative level.

Despite this, the application of methods and results from more quantitative sciences is slowly making inroads. Most sediment is deposited after transport in water or air, so there is an obvious connection between sedimentary geology and fluid mechanics. This is emphasized early on in Leeder's short introduction to fluid mechanics followed by material on sediment transport and various sedimentologically relevant flow types. He is clearly fascinated with turbulence, and this topic is presented in great detail, albeit somewhat descriptively. He includes nice treatments of important topics, such as rotating flows, not typically found in introductory fluids courses. Equally impressive is the attention paid to putting geochemical aspects of sediments, especially carbonates, on a fundamental footing. Most sedimentary geology texts of recent years pay lip-service to these topics; Leeder's treatments are substantial.

A few things might be improved in subsequent editions. The quantitative treatment suffers from an uncritical acceptance of some of the more dubious ideas of R. A. Bagnold, whose undisputed status as a giant in sediment mechanics does not make him infallible. Climate is only cursorily

New journals

This year, *Nature's* annual new journals review supplement will appear in the issue of 21 September. Publishers and learned societies are invited to submit journals (including electronic journals) for review, taking note of the following criteria:

- Journals must have first appeared during or after June 1998, and paper journals must have published at least four separate numbers by the end of May 2000.
- Journals covering any aspect of science are eligible, although those dealing with clinical medicine and pure mathematics are excluded, as are newsletters and publications of abstracts.
- Frequency of publication must be at least three times a year.
- The main language must be English.
- Deadline for submission is 5 June.

Please send at least four different issues (the first, the most recent and any two others) of each eligible title, or the URL and access information for electronic journals, together with full details of subscription rates, to: Isobel Flanagan, *Nature*, Porters South, Crinan Street, London N1 9XW, UK. Tel: +44 (0)20 7843 4542. e-mail: i.flanagan@nature.com

connected to sedimentology at the end of the chapter. Also missing is much on the very active subfield of organisms and sediments, including the study of 'trace' (behavioural) fossils and their considerable impact on the stratigraphic record. Sedimentary basins, in whose dynamics lies the strongest connection between sedimentary geology and tectonics, are covered very briefly. There is nothing on seismic or sequence stratigraphy, or on stratigraphic modelling — a shame, as quantitative stratigraphic modelling is perhaps the one area that spans the full range from turbulence to tectonics.

On the whole, though, what is missing is far less important than what is there. Nothing else approaches this book in scope, depth or liveliness of presentation. Sedimentary geology today is like a medieval bazaar in which scientists speaking languages as disparate as particle physics and geography, civil engineering and evolutionary biology, are all contributing to the mix. Very few people could have drawn as much of this work together as coherently as Mike Leeder has. Anyone who would translate the Book of Sediments should start here. ■

Chris Paola is in the Department of Geology and Geophysics and St Anthony Falls Laboratory, University of Minnesota, 3rd Avenue SE and Mississippi Rivers, Minneapolis, Minnesota 55455-0219, USA.

On the scent of the sixth sense

Jacobson's Organ: And the Remarkable Nature of Smell

by Lyall Watson

W. W. Norton: 2000. 272 pp. \$24.95

Michael Stoddart

Jacobson's organ is an anatomical structure lying in the nasal septa of a wide range of animals. It is named after Ludwig Levin Jacobson, an eighteenth-century Danish army surgical officer with a penchant for comparative anatomy. His discovery was reported to the world by Baron Georges Cuvier, of the Musée d'Histoire Naturelle in Paris. "The organ consists of a long narrow bag of gland-like substance," he told his students in 1811, "surrounded by a cartilaginous case of the same form, located on the floor of the nasal cavity, on each side, very near the ridge on which rests the inferior border of the cartilaginous portion of the nasal septum ... M. Jacobson has examined this organ in various animals which possess it ... we find it most developed in rodents, next in ruminants. The carnivores have less, and in monkeys it becomes so small that we are prepared to see it vanish completely in man."

Although Jacobson noted well-developed nerves passing dorsally from the organ, he concluded that the organ was secretory. But he remained puzzled as to

why it should provide secretion to the nose and mouth, with which it is connected via the nasopalatine canal.

We now know that Jacobson's organ functions as a secondary organ of smell, its distinct nerves passing not to the olfactory bulb of the brain but to an accessory olfactory bulb. With connections to the regions of the brain that control sexual behaviour and functioning, Jacobson's organ is used by most species to allow the animal to determine the state of sexual readiness in potential mates. The stallion's dramatic flaring of the lips after tasting the mare's urine is a well-known example of how the lumen of the organ becomes suffused with sexual pheromones.

Lyall Watson's easy, flowing text picks its way through the comparative anatomy, biochemistry and evolutionary biology of olfactory communication in the animal kingdom in a most readable way. Jacobson's organ and its role in animal behaviour are described in some detail. Interspersed are vignettes about Linnaeus's classification of odours, together with literary references and anecdotes describing some of the enigmas that constitute human olfaction. Watson reminds us of the remarkable olfactory powers of Helen Keller, the blind and deaf American author who could recognize people by their odour; the lustful recall of Napoleon Bonaparte's pleasure in the smell of Josephine's body; and the *mémoire involontaire* described by Marcel Proust, when the aroma of a dunked madeleine biscuit unleashed a flood of childhood memories.

Without realizing it, the reader is invited

to consider what Watson calls "the sixth sense" as a category of perception dependent upon Jacobson's organ for its sensory input. "We have Jacobson's Organ," he writes. "It is present in almost every human nose so far examined. And it appears to be functional: a miniprobe inserted into the appropriate pits on either side of the nasal septum registers a measurable potential in the presence of steroidal pheromones." He cites the powers of the *ombiasy*, or local herbal healers, of Madagascar: a healer sniffs out a solution to a particular patient's condition "by wandering with an open mind, until something stops him, until a particular plant catches his attention, and gives a whole new slant to the idea of an elective procedure, by offering itself as the remedy". He adds that we should approach the plant world with flared lips "in a way that opens up the duct to Jacobson's Organ".

What Watson omits to inform us is that the presence of a functional Jacobson's organ in humans is highly contentious. Despite almost two centuries of comparative anatomical research since Jacobson's day, there remains no unequivocal evidence for the presence of a functioning Jacobson's organ in any Old World monkey, ape or human. True, one study has reported the presence of pits on either side of the nasal septum in most of a sample of 1,000 people, but another, involving 1,842 adults, found paired nasal pits in just 13 per cent of them. Another has reported apparent evoked potentials from these pits when they were bathed in putative human pheromones, but the possibility that this was the result of stimulation of the trigeminal nerve was not addressed. Some studies have reported the presence of bipolar receptor cells in the organ's lining, similar to those found in the olfactory mucosa, whereas others have explicitly failed to find such receptors. An accessory olfactory bulb — a prerequisite for the functioning of Jacobson's organ — has not been described for any Old World primate, including humans.

To the critical scientist, the case for the existence of a functional Jacobson's organ has yet to be made. Much of the support for its existence in humans has come from a company that markets pheromone-enriched perfumes, allegedly acting via the organ and its pathways to the sexual centres of the brain. I await independent corroboration. Maybe Watson's book is well ahead of its time; even he thinks it might be, as he reveals by admitting that discourse on the sixth sense is speculative.

This is not a book for the serious scientist. Its uncritical approach will only fuel further speculation about the supposed supernatural powers of the sense of smell. ■

Michael Stoddart is in the Australian Antarctic Division, Channel Highway, Kingston, Tasmania 7050, Australia.

A whiff of the past

The Ancient Egyptians valued perfumes and cosmetics highly, and the Egyptian climate, with its regular supply of water and abundant sunshine, was well suited to producing floral fragrances. The priestess Ihat, seen here sniffing a lotus, Egypt's most valued aromatic

flower, is featured in *Sacred Luxuries: Fragrance, Aromatherapy and Cosmetics in Ancient Egypt* by Egyptologist Lise Manniche, with photographs by Werner Forman

(Opus, £24.95, \$30.95).



Horse power

The millennium of the horse began with a whimper, but went out with a bang.

Vaclav Smil

Horses make superior draught animals. Unlike cattle, their front ends are heavier than their rears (by a ratio of 3:2), an advantage in inertial motion. They can stand without engaging muscles — and hence without additional metabolic cost — by ‘locking’ their legs with suspensory ligaments and a pair of tendons. And they grow larger and live longer than cattle, and have greater endurance.

Medieval Europe adopted collar harnesses and iron horseshoes just before the end of the first millennium. Ancient throat-and-girth harnesses choked the animal during heavy exertion, and breastband harnesses precluded the optimal transfer of power; fitted and padded collar gave a desirably low traction angle while deploying powerful breast and shoulder muscles without restricting breathing. Iron horseshoes saved hooves from excessive wear and improved traction, and swingletrees, attached to traces to equalize the strain resulting from uneven pulling, made it possible to harness an even or odd number of horses.

By 1000, horses were thus mechanically ready for a bigger role in medieval society — and yet it took them several centuries to displace oxen as the principal draught animals. The Domesday count of 1086 shows that a mere 5 per cent of all demesne draught animals, and about a third of the total on English peasant holdings, were horses. By 1300 they made up nearly half of all draught animals owned by peasants. But it was more than 200 years before they became dominant. The size of the animals, their nutrition, and the ploughs they pulled explain why.

The body weight, and hence power, of draught horses only began to rise after several centuries of breeding the heavy war animals needed to carry armoured knights. And more powerful horses needed at least some concentrates, such as cereal or legume grains, in their diets — not just the roughages (grasses, straws) fed to cattle. Farming had to be intensified to provide this feed for both people and animals. This was a slow process: in England, crop harvests were stagnant for centuries, and average yields only took off in the late eighteenth century.

But even larger, well-fed horses had a tough time with wooden ploughs. The ploughs’ heavy soles, wheels and mould-boards generated enormous friction, particularly in wet soils, and without a smooth, curved fitting between the share and the flat mould-board, they were constantly clogged

with soil and weeds. Iron mould-boards only crossed from China to Europe in the seventeenth century, and it was not until the rise of the modern steel industry in the mid-nineteenth century that smooth, curved, steel ploughshares replaced cast-iron shares.

Increased body size, improved feed quality and cheap steel transformed draught horses into unrivalled energizers of the early modern world’s agriculture. Horses were also essential during the initial stages of industrialization, powering road and canal transport, and turning whims in mining, food processing and numerous manufactures before they were displaced by steam engines.

Heavier draught animals made it possible to cultivate the existing fields more frequently and to convert forests and grasslands to new fields more easily, and freed labour for other activities. Most notably, the post-1850 ploughing-up of America’s enormous grasslands could not have proceeded so quickly with oxen. The heaviest nineteenth-century breeds — French percherons, English shires, German rheinlanders — approached and

even topped 17 hands and weighed around one tonne. During brief exertions such horses could develop about three horsepower, and they could work steadily at rates well surpassing one horsepower (745 W). These animals needed a lot of feed, but their energy advantage was indisputable: a horse eating 4 kg of oats a day consumed about the same amount of grain as six people — but its power could supplant that of at least ten strong men.

By the 1890s, American horses, working in teams of more than 30 animals, pulled not just multishare ploughs but also the first grain combines. Such large numbers of animals needed abundant farmland: when the number of farm horses in the United States peaked at 21 million in 1919, at least 20 per cent of the country’s farmland was needed to cultivate their feed. By that time internal-combustion engines and electric motors had almost totally displaced draught horses in the cities of industrial countries.

Their eventual replacement in Western farming was inevitable: even the small internal-combustion engines mounted on the first tractors could replace at least ten horses — and claimed no land. By the end of 1960, when three million horses were still left on American farms, the US Department of Agriculture stopped counting draught animals. The millennium of gradually expanding horse power that was so instrumental in the rise of the West ended in a rapid retreat. ■

Vaclav Smil is in the Department of Geography, University of Manitoba, Winnipeg R3T 2N2, Canada.

It took several centuries for horses to displace oxen as the main draught animals.



Pulling power: better harnesses and ploughs put horses’ efforts to more efficient use.

Disextinction, Inc.

For his biology project, Jason decided to resurrect the passenger pigeon.

Kathryn Cramer

“Daddy’s!” shouted Erika, my bright-eyed granddaughter, when she saw the new bears at the zoo. So many of the animals — moas, dodos, thylacines — have a fluorescent green gene with the copyright of my son Jason’s company, Disextinction, Inc. Today we saw the newest: atlas bears, absent since the nineteenth century.

In 1999, when Jason was a toddler learning to talk, he asked to see the dodo birds at the zoo. “Go zoo see dodo birds! Go zoo see Barbary lions!” He was an animal kid. Selecting library books, he chose those with realistic faces of animals on the cover.

I had just read him *Gone Forever! An Alphabet of Extinct Animals*. I exchanged solemn glances with the children’s librarian and explained: extinction is for ever. There were never going to be any more. These animals were all gone.

“All gone,” he echoed earnestly. “So sad.”

For days he pestered me: “Want quaggas! Want great auk!” I searched the web for animals he craved. I searched on Barbary lion, allegedly extinct since the 1920s. The preferred lions of ancient Roman arenas, they were bigger than the lions in the zoo. I don’t know how to convey my shock when the search engine delivered links to recent photographs of Barbary lions rescued from an abandoned circus in Zanzibar.

I found a photograph of the last quagga, striped from head to midriff, which died in Amsterdam in the 1870s. I also found the Quagga Rebreeding Project, with recent photographs of rebred quaggas grazing on the grounds of a particle accelerator in South Africa. A plains zebra subspecies, the quagga was rebred from the living population.

As I dug deeper, the stories got weirder. Schoolboys in Hastings, New Zealand, held an academic conference on the possibility of cloning their school mascot, the huia bird, driven extinct by a 1920s hat craze. The assembled scientists set out to clone the huia.

In Thailand, there was a project to clone a magnificent 100-year-old preserved white elephant for the king. When the Japanese woolly mammoth group got its tissue samples to reproduce, the living cells were to be sent to the white-elephant cloners.

I opened Jason a brokerage account and put in his and my Christmas money. My thought was to buy him stock in Disney or Mattel. Then I had a better idea. A company called Geron was down because they had



bought Roslin Labs — the Roslin that cloned Dolly the sheep. Bingo!

I bought him a hundred shares. I put more money into what I was thinking of as his college fund and bought other interesting biotech stocks: Abgenix, Affymetrix, Maxygen. Wall Street thought these stocks were interesting too. My son became quite well-to-do for a toddler.

When Jason was three, we drove to Florida to see his first Barbary lion. When he was four, we went to Africa to see quaggas. When he was five, we went to New Zealand to see fuzzy little huia chicks, born of magpies, and a herd of Enderby cattle, a breed that can live on a diet of seaweed, rebred from a single ageing cow and some frozen bull semen. When he was eight, we flew to Tokyo to see baby woolly mammoths. When he was 12, we visited the Barbary Lion Refuge in the mountains of Morocco.

At 14, he decided to do some cloning of his own. For his biology semester project he chose the passenger pigeon. Because of the boom in livestock cloning, all the services he needed could be located through Enron Online’s Biotech Market. Once he talked a tissue sample out of the Smithsonian, all he had to do was order services over the web and send and receive FedExes. Living chicks arrived by the semester’s end.

Suddenly Jason was famous as the kid who’d cloned the passenger pigeon. He’d spent a third of his ‘college fund’ on eight passenger pigeon chicks. I was amazed and appalled. Then he got tissue samples from the Nationaal Natuurhistorisch Museum Leiden and spent the rest of his college fund cloning the great auk and, peculiarly, the quagga.

He sold his great auk chicks at a significant profit. The quagga colt he donated to the quagga rebreeding project. He incorporated, calling his company Disextinction, Inc. “We’re at the beginning of a mass extinction. I’ve got a service the world needs,” he said. When he was 19, Disextinction went public. By that time he was too busy for college.

When I see Erika cuddled up to Javanese tiger cubs in front of Jason’s tank of exotic cichlids, I can’t help but wonder what else Jason has added to their genetic make-up, how disextinct species differ from the originals.

Yesterday Disney bought the company. When Erika heard the news, her eyes grew large! She is so pleased.

Kathryn Cramer has co-edited the science fiction anthologies *The Ascent of Wonder (Orbit)* and *Spirits of Christmas (Tor)*. She lives in Pleasantville, New York.

Designed to dissolve

Walter Leitner

Many solvents are unpleasant but essential industrial chemicals. Supercritical carbon dioxide would be a viable 'green' alternative but its use has been restricted by its limited solvent power. This is about to change.

Compressed carbon dioxide is often promoted as an environmentally friendly solvent having useful properties for a wide range of technical and chemical processes¹⁻³. A growing number of commercial operations and pilot activities illustrate that CO₂ — especially in its supercritical state — can combine ecological and economic benefits. But the poor solubility of many interesting target substances places a severe restriction on the widespread use of CO₂ as a solvent. In many cases, either unrealistically high pressures or expensive 'CO₂-philic' materials (substances with a high affinity for CO₂ solutions at lower pressures) are needed. On page 165 of this issue, Eric Beckman and his colleagues⁴ from the University of Pittsburgh introduce a new strategy for designing CO₂-philic materials from readily available components, which overcomes some of the oldest prejudices about the solvent properties of CO₂.

If you are sipping a cup of decaffeinated coffee while reading this journal, it is most likely that the caffeine was extracted from the coffee beans using CO₂ as a solvent. In case you prefer to read your literature over a pint of lager in the evening, there is again a fair chance that some of its hops' aroma was produced with this technology. Carbon dioxide is used in these processes under 'supercritical' conditions, that is, at pressures and temperatures beyond the critical values (Fig. 1). No distinct gas or liquid phase can exist above the critical point, and the supercritical phase has a unique combination of properties from both states. At liquid-like densities supercritical carbon dioxide (scCO₂) exhibits low viscosity and high diffusion rates, just like a gas. Owing to the high compressibility of the supercritical phase, its solvent properties can be varied by small changes in temperature and pressure.

In the late 1960s, Kurt Zosel at the Max Planck Institute for Coal Research in Mülheim recognized the potential of non-toxic and environmentally benign scCO₂ for natural product extraction⁵. The advantage of scCO₂ over conventional extraction with liquid solvents is that the supercritical fluid penetrates the material efficiently and can be removed after purification, without leaving a trace, simply by lowering the pressure below the critical value. Zosel developed extraction methods, now widely used in

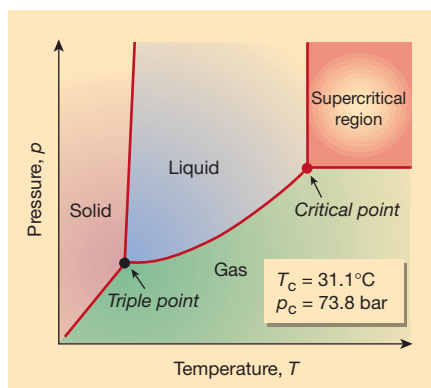


Figure 1 Phase diagram for carbon dioxide. Beyond a specific temperature and pressure (the critical point) carbon dioxide becomes a supercritical fluid, a state that is neither a gas nor a liquid, but has properties of both.

industry, by exploiting the fact that scCO₂ is a rather choosy solvent: it will selectively strip off caffeine, leaving behind those substances responsible for the coffee's flavour.

The specific properties of scCO₂ make it an interesting 'green' replacement for organic solvents, which are often less than ideal owing to their acute toxicity, ecological hazards or difficulties with disposal and recycling. Supercritical CO₂ may find applications in areas as diverse as the dyeing and cleaning of fibres and textiles, polymerization and polymer processing, purification and crystallization of pharmaceuticals, and

— last but not least — as a reaction medium for chemical synthesis. Supercritical CO₂ can be handled safely in standard high-pressure equipment on a laboratory or industrial scale. But the limited ability of CO₂ to dissolve polar or non-volatile compounds represents a major drawback in many processes, because key components will often fail to form homogeneous solutions under practical conditions. The design of CO₂-philic solubilizers to aid this process is therefore paramount in these areas.

A general design principle for CO₂-compatible materials is to tether the desired substances to one or more solubilizers with a very high affinity for CO₂ (Fig. 2). The 'CO₂-phobic' part can be a functional unit such as a dyestuff, a reagent or a catalyst, or it might be designed to interact with other CO₂-insoluble molecules, giving the whole ensemble the function of a surfactant. The key problem is then reduced to the design of highly effective and economically feasible solubilizers.

Until now, the most widely used CO₂-philic solubilizers have been polysiloxanes and fluorocarbons. The latter have been particularly effective, allowing, for example, the design of surfactants to disperse polymers⁶ or even (highly polar) water⁷ in CO₂, the synthesis of ligands for metal extraction⁸ and homogeneous catalysis⁹, or the use of otherwise insoluble organic reagents¹⁰. The Beckman group⁴ has now synthesized a non-fluorous but still highly CO₂-philic

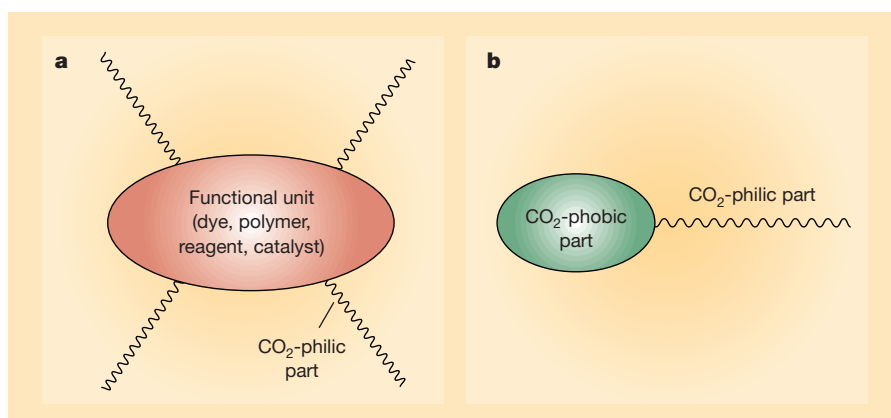


Figure 2 The design of CO₂-philic materials. Solubilizers with high affinity for CO₂ solutions (wavy lines) are either a, attached directly to functional units or b, used to generate CO₂-soluble surfactants with emulsifying or dispersing properties. The most widely used solubilizers are fluorinated hydrocarbon chains, which are expensive and must be incorporated using elaborate synthetic procedures. Beckman *et al.*⁴ describe a new strategy for designing CO₂-soluble materials from cheap and readily accessible feedstocks that will increase the practical applications of supercritical CO₂.

polymer, whose solubility in scCO_2 results from a judicious design. Their success is based on a rational analysis of the entropy (structure) and enthalpy (interactions) contributing to the solubility of these polymers in CO_2 .

The target compound of the Beckman group's study was a copolymer, created by linking together two chemically distinct repeating units (here, polyether and polycarbonate groups). Such copolymers can be easily generated using an aluminium catalyst to react inexpensive propylene oxide ($\text{C}_3\text{H}_6\text{O}$) with — ironically — CO_2 itself. The polyether skeleton of these copolymers is highly flexible and has only weak polymer–polymer interactions. Introducing the carbonate groups enhances this flexibility, and therefore the entropy of mixing, even further. At the same time, a favourable interaction of the carbonyl group with CO_2 may increase the enthalpy of mixing, thereby also improving the solubility of these compounds in CO_2 .

The success of this new design principle is convincingly demonstrated by the creation of polymers consisting of 25–250 repeat units of propylene oxide or ethylene oxide, and containing 15–40% carbonate linkages. Although prevailing opinion suggested that hydrocarbon polymers of even moderate molecular weights are practically insoluble in scCO_2 at medium pressures (less than 300 bar), these new materials show excellent solubilities exceeding even those of fluorinated polyethers of comparable molecular mass. Furthermore, the ability of these copolymers to dissolve in scCO_2 is dramatically increased compared to homopolymers formed from either 100% polycarbonate or 100% polyether.

Being the first highly CO_2 -soluble hydrocarbon polymers, these new materials are far more than just curiosities. The aluminium catalyst remains attached to the growing polymer chain during the copolymerization process, providing an obvious target for the connection of functional groups to the CO_2 -philic material. This strategy could be confirmed by modification of the new copolymers with carboxylate endgroups, leading to the first non-fluorinated surfactants that can stabilize emulsions of scCO_2 and water. The opportunity to fix catalysts or reagents through similar routes is obvious, demonstrating the potential of Beckman and colleagues' copolymers as a new class of solubilizers for scCO_2 . These new structures form a welcome and long-awaited addition to the general toolbox shown in Fig. 2, and expand both our practical opportunities and our understanding of how to design CO_2 -philic materials. ■

Walter Leitner is at the Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, D-45470 Mülheim an der Ruhr, Germany.
e-mail: leitner@mpi-muelheim.mpg.de

1. McHugh, M. & Krukonis, V. J. *Supercritical Fluid Extraction* 2nd edn (Butterworth-Heinemann, Boston, 1994).
2. Jessop, P. G. & Leitner, W. (eds) *Chemical Synthesis Using Supercritical Fluids* (Wiley-VCH, Weinheim, 1999).
3. Noyori, R. (guest ed.) *Chem. Rev.* **99**, issue 2 (1999).
4. Sarbu, T., Styranec, T. & Beckman, E. J. *Nature* **405**, 165–168 (2000).
5. Zosel, K. *Angew. Chem. Int. Edn Engl.* **17**, 702–709 (1978).

6. DeSimone, J. M. *et al. Science* **265**, 356–359 (1994).
7. Johnston, K. P. *et al. Science* **271**, 624–626 (1996).
8. Laintz, K. E., Wai, C. M., Yonker, C. R. & Smith, R. D. *Anal. Chem.* **64**, 2875–2878 (1992).
9. Kainz, S., Koch, D., Baumann, W. & Leitner, W. *Angew. Chem. Int. Edn Engl.* **36**, 1628–1630 (1997).
10. Hadida, S., Super, M. S., Beckman, E. J. & Curran, D. P. *J. Am. Chem. Soc.* **119**, 7406–7407 (1997).

Hearing

Tuning in with motor proteins

Matthew Holley

Most animals use hair cells in the ear to detect sound. Each hair cell has a highly organized bundle of filaments at its apex that can change sound-induced vibrations into electrical signals. The mammalian ear also has a more specialized cell population (Fig. 1), which detects and converts sound in the normal way but then converts the electrical signals into changes in cell length. This extraordinary response appears to be an essential component of a mechanical tuning mechanism that is unique to the mammalian ear. Zheng and colleagues¹, writing on page 149 of this issue, have now identified the motor protein that drives this response. Their work opens up opportunities that will help us to understand our own hearing and to harness the forces of one of the smallest and fastest biological motors.

The sensory part of the mammalian auditory organ is housed within a coiled tube called the cochlea (Fig. 1a). The hair cells are arranged with exquisite precision within the organ of Corti, a cellular layer that lies on a long, narrow membrane called the basilar membrane (Fig. 1b). The width and stiffness of the basilar membrane change progressively along the cochlear coil, so that it vibrates best to high- and low-frequency sounds at its base and apex, respectively. This forms a map of incoming sound frequencies that is transmitted to an associated array of auditory nerve fibres. The interaction between vibrations of the basilar membrane and the sensory hair bundles is extremely complex (Fig. 1b). One particularly striking feature of the organ of Corti is that it contains two distinct populations of hair cells. There is a single row of goblet-shaped, conventional, inner hair cells and several rows of cylindrical outer hair cells (Fig. 1b). Why do mammals need these two types?

The physical properties of the basilar membrane alone cannot account for the known sensitivity and frequency selectivity of the living ear. Sellick *et al.*² showed that it is very finely tuned, and Davis³ subsequently described the need for a 'cochlear amplifier'. The idea that the outer hair cells might be the amplifiers was dramatically reinforced by the finding⁴ that these cells can change length

rapidly when electrically stimulated. The outer hair cells have since been subjected to a whole range of experimental manipulations, leading to a widely accepted explanation for the cellular mechanism⁵. In this theory, arrays of densely packed, voltage-dependent motor proteins in the membrane of the outer hair cell force the cell's surface area to change. The cell is constrained by a two-dimensional protein skeleton (Fig. 1c), so that changes in its surface area are expressed as changes in cell length.

Nonetheless, the function of the outer hair cells remains unclear — studies of isolated cells tell us as much about cochlear mechanics as lighting an isolated, fuel-soaked piston tells us about the internal combustion engine. But the identification of the gene encoding the motor protein¹ provides a powerful tool with which to address this problem.

The method adopted by Zheng *et al.*¹ to identify that motor protein is conceptually elegant, exploiting the two types of hair cell to identify genes specifically expressed in each. The apparent abundance of motor protein in the outer hair cell indicated that these cells may contain high levels of the messenger RNA encoding that protein. The level of mRNA is a measure of how highly the relevant gene is expressed. With 1,000 inner and outer hair cells as a source of mRNA, the authors made expression libraries and compared them by subtractive hybridization, reasoning that a gene that was expressed exclusively in the outer hair cell might encode the motor protein responsible for the changes in cell shape. With more than 20,000 genes expressed in a single cell, these techniques usually depend on a good deal of luck. In this case, however, the close relationship between the two cell types was a clear advantage. One of the candidate genes selected by Zheng *et al.* happened to encode a protein that is closely related to members of a family of anion/sulphate transporters.

The acid test of a candidate motor is whether it can make a cell 'jump' like an outer hair cell. So Zheng *et al.* put their candidate gene into cultured kidney cells. It would seem unreasonable to expect a single gene in

an alien cell to express a functional membrane protein in sufficient density to recreate the motile response of a mature hair cell. So the authors looked for gating currents and voltage-dependent changes in cell capacitance — the electrical signature of the motor protein in normal outer hair cells. Gating currents are small, transient currents caused by the displacement of charged structural elements of a protein as it changes shape within a membrane. With a high density of protein, these charge displacements are associated with a voltage-dependent change in cell capacitance.

There must have been a rare sense of excitement when Zheng *et al.* started to analyse their experimental kidney cells. They not only saw a convincing electrical signature, but also recorded changes in cell shape. Critics may argue that any voltage-sensitive membrane protein expressed at sufficient

density will generate the same effect. But cells transfected with pendrin, a similar transporter molecule, did not respond in this way. Furthermore, the response was at least partially inhibited by salicylate, which inhibits length changes in outer hair cells. The candidate gene, since christened *Prestin*, is also expressed specifically in cochlear outer hair cells. Its expression during development coincides with appearance of the membrane protein and with activation of cell length changes.

The work of Zheng *et al.*¹ will open up many research avenues. The predicted amino-acid sequence encoded by the *Prestin* gene yields little in terms of the structure of the protein, but the secrets of this new molecular motor should be revealed by mutational analysis both *in vivo* and *in vitro*. Modification or deletion of the gene in adult animals offers a new and more powerful way

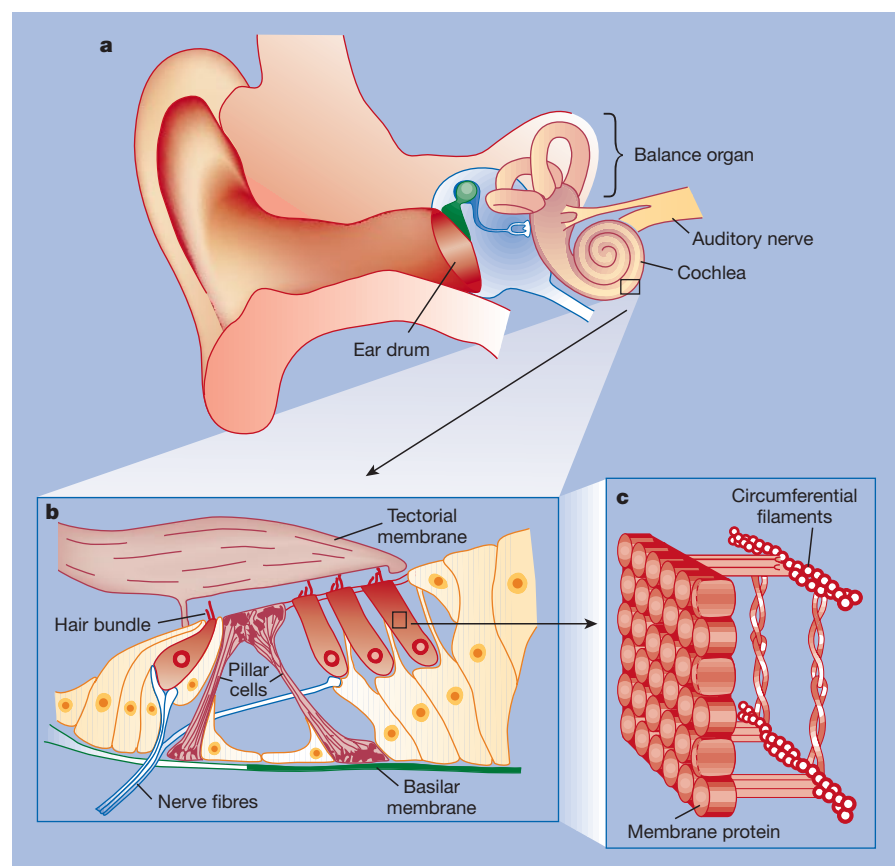


Figure 1 How we hear. **a**, The cochlea. Sound enters the outer ear canal, vibrates the eardrum and is transmitted to the cochlea by the middle ear bones. **b**, A transverse section through the organ of Corti. Inside the cochlea, the organ of Corti sits upon the basilar membrane. Its surface is covered by another membranous structure, the tectorial membrane. Vibration of the basilar membrane leads to lateral displacements of the mechanosensory bundles on the hair cells (coloured orange). Inner hair cells form a row to one side (the left side here) of two rows of specialized supporting cells (pillar cells); outer hair cells form several rows on the other side. Inner hair cells are connected mainly to afferent nerve fibres and so are responsible for sending sensory information to the brain. Outer hair cells are connected to relatively few afferent fibres and are located above the most flexible region of the basilar membrane, where they can more easily influence its mechanical responses. **c**, The plasma membrane of outer hair cells contains a high density of membrane proteins embedded within the lipid layer. It is connected to relatively stiff, circumferential filaments that are crosslinked by thinner, more compliant filaments. This submembrane skeleton helps to convert changes of membrane surface area into changes in cell length.



100 YEARS AGO

The Paris correspondent of the *Chemist and Druggist* states that science is represented at the Salon by several portraits of average merit. The best is that of Dr. Vaillard, head army surgeon and professor at the Val de Grace Military Hospital, where he is known to two or three generations of army pharmacists who have followed his lectures. Dr. Vaillard is of middle age, and is shown standing, in regimental dress, with the Cross of the Legion of Honour on his tunic. His left hand is leaning on a laboratory-bench, on which are a microscope and a variety of analytical appliances... One would like to see more of this class of picture, but must suppose artists find no market for them.

From *Nature* 10 May 1900.

50 YEARS AGO

In 1940 the first isotopes of the elements 93 (neptunium) and 94 (plutonium) were produced, and four years later the artificial creation of the elements 95 and 96 (americium and curium) was accomplished... Now from the Radiation Laboratory of the University of California comes the news that within the past few months the next two transuranic elements have been made and identified. The first announcement, released early in January of this year, concerned element 97. S. G. Thompson, A. Ghiorso and G. T. Seaborg, aided by a number of other physicists and chemists working under the auspices of the U.S. Atomic Energy Commission, had bombarded the isotope of element 95 (americium), of atomic weight 241, with 35-MeV. helium ions accelerated in the California 60-in. cyclotron and thereby obtained a nuclide with a half-life of 4.6 hr. and probable atomic weight 243. It decays by electron capture, with approximately 0.1 per cent α -branching; there seem to be three α -particle groups, that of 6.72 MeV. having the highest energy. The production of element 98, which has just been reported by the same three authors and K. Street, jun., was achieved in an analogous way. By irradiating the isotope of element 96 (curium) of atomic weight 242, in the same cyclotron with helium ions of the same energy, a new nuclide was formed which decays with a half-life of about 45 min., emitting α -particles of energy 7.1 MeV. The possible and, for theoretical reasons, likely decay by electron capture has not yet been observed. The atomic weight should be 244.

From *Nature* 13 May 1950.

of studying the role of outer hair cells in hearing. Given that outer hair cells can operate at frequencies of at least 80 kHz (ref. 6), *Prestin* may also have a bright future in biological nanotechnology. Finally, I can't help wondering about some of the other spots on the research team's subtraction blots. One of them ought to be the sugar transporter GLUT5, proposed by Géléc *et al.*⁷ as a candidate for the motor. It might be worth trying that protein in a kidney cell, too. ■

Matthew Holley is in the Department of Physiology, School of Medical Sciences, University of Bristol,

University Walk, Bristol BS8 1TD, UK.

e-mail: m.c.holley@bristol.ac.uk

1. Zheng, J. *et al.* *Nature* **405**, 149–155 (2000).
2. Sellick, P. M., Patuzzi, R. B. & Johnstone, B. M. *J. Acoust. Soc. Am.* **72**, 131–141 (1982).
3. Davis, H. *Hearing Res.* **9**, 79–90 (1983).
4. Brownell, W. E., Bader, C. R., Bertrand, D. & de Ribaupierre, Y. *Science* **227**, 194–196 (1985).
5. Holley, M. C. in *The Cochlea: Springer Handbook of Auditory Research* Vol. 8 (eds Dallos, P., Popper, A. N. & Fay, R. R.) 386–434 (Springer, New York, 1996).
6. Frank, G., Hemmert, W. & Gummer, A. W. *Proc. Natl Acad. Sci. USA* **96**, 4420–4425 (1999).
7. Géléc, G. S. G., Casalotti, S. O., Forge, A. & Ashmore, J. F. *Nature Neurosci.* **2**, 713–719 (1999).

Cosmology

The dark side of distortion

Max Tegmark

If you see your reflection in a spoon, your face will look strangely elongated because of the way the light rays are bent. Einstein predicted that gravity could also bend light, so you should in principle see similar image distortions when looking out into space. In astronomy this effect is known as gravitational lensing, and many examples have now been found of such distortions. For instance, a spherical galaxy located far behind a massive galaxy cluster tends to look elongated. So gravitational lensing offers a useful way of measuring the density and distribution of matter between distant galaxies and us, even if the matter itself cannot be seen. The more mass that is present, the larger the distortions should be.

A long-standing goal in cosmology has been the detection of lensing, not just from compact objects such as stars, galaxies or galaxy clusters, but also from the most

elusive matter in the Universe. This is the 'dark matter' that appears to fill intergalactic space and to weigh about ten times as much as all the gas and stars combined. The gravitational effect of this dark matter has now been observed by Wittman *et al.*¹ (reported on page 143 of this issue), and by three other groups announcing similar results^{2–4}. Tony Tyson and his group at Bell Labs, of which David Wittman is a member, have been searching for gravitational lensing from large-scale structure in the Universe for 15 years. Why did it take them and the other teams so long to detect it?

The first problem is that this lensing effect is weak. The distorted image of a perfectly spherical galaxy becomes only slightly elliptical, with the major and minor axes differing by about 1%. In contrast, real galaxies are often far from spherical, and can easily appear twice as wide in one direction

as in another. Because we have no independent way of knowing whether a galaxy is intrinsically elliptical, there is no way of determining the extent to which any one galaxy is distorted by dark matter. Fortunately, this problem can be overcome statistically. Truly elliptical galaxies are expected to be randomly orientated, because even apparently close galaxies are usually at vastly different distances from us and therefore from each other. But gravitational lensing will, on average, tend to elongate galaxies in the same sense if they lie behind the same dark matter concentrations, as illustrated in an exaggerated way in Fig. 1. Wittman *et al.*¹ use about 145,000 galaxies in their analysis, and find their elongations to be correlated in exactly the way predicted by gravitational lensing.

The most difficult challenge in the weak-lensing game is making sure that these correlations are real, because a number of potential systematic errors can cause similar effects. For instance, most real-world telescopes cause slight image distortions, which can be correlated across the entire field of view. Wittman *et al.*¹ made their observations with a custom-made instrument called the Big Throughput Camera on the Cerro Tololo telescope in Chile, and used a battery of techniques to quantify and control such effects. For instance, stars in our own Galaxy should be free of true lensing effects, and are used by Wittman *et al.* to quantify and remove residual image distortions.

The main consequence of the present work may not be the measurements *per se*, which are of the magnitude expected from standard cosmological models. Rather, the detections by these four groups^{1–4} could be the first shots of a revolution in our ability to measure dark matter. The situation is rather analogous to that of the cosmic microwave background (CMB) — the faint glow of radiation generated by the Big Bang — in which tiny temperature variations were discovered by the COBE satellite in 1992. Although it had been known for many years that measurements of the CMB and gravitational lensing could be used to probe the properties of dark matter — through well-understood gravitational effects — many sceptics doubted that the daunting experimental challenges could be overcome. The COBE observations triggered an explosive growth in that field, which entered an era of true precision measurements with the announcement last month of results from the Boomerang experiment⁵. The next decade might hold similar promise for weak-lensing observations, with a second generation of surveys covering much larger areas of sky and using even better technology.

Cosmology is different from other areas of physics in that the numerical values of the parameters that specify the current 'standard model' are poorly known, some not even to

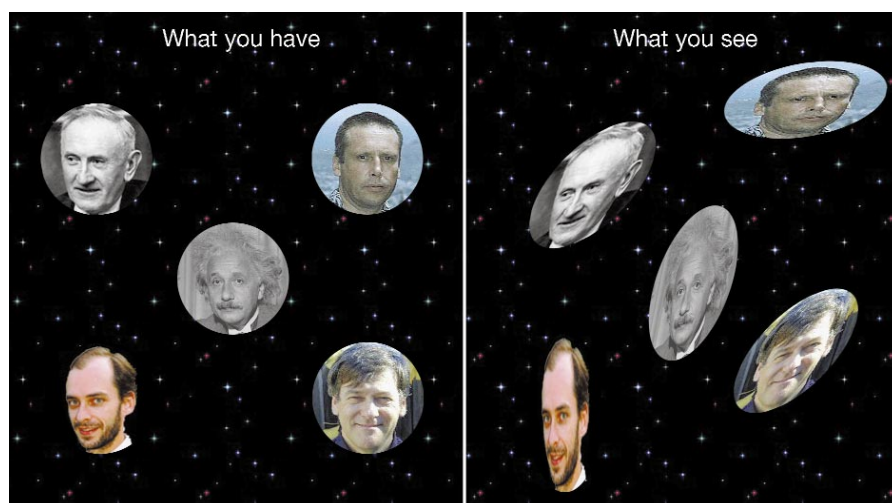


Figure 1 Gravitational lensing distorts space so that spherical images become elliptical. Moreover, the directions in which nearby images become elongated are nearly parallel, as illustrated by some early theorists in the field (clockwise from top left, Fritz Zwicky, Nick Kaiser, Roger Blandford and Jordi Miralda-Escude, with Albert Einstein in the centre). Wittman *et al.*¹ use similarly distorted images of distant galaxies to estimate the amount of dark matter in the Universe.

within a factor of two. These cosmological parameters include the amounts of ordinary matter and dark matter in the Universe, the curvature of space and the amplitudes of various types of density fluctuations that emerged during the first split second after the Big Bang. Moreover, we still don't know what this dark matter is — current models suggest it might be dominated by so-called vacuum energy and exotic particles.

But this low-precision era of cosmology could soon be over thanks to an avalanche of new high-quality data from areas such as the CMB, the clustering and motion of galaxies, distant supernovae, gas clouds absorbing light from quasars and various forms of gravitational lensing. Each type of observation tells us about certain combinations of the parameters, so no single approach will

measure them all. Rather, the key will be to combine different types of measurements, both to obtain true precision measurements of the cosmological parameters and to provide independent crosschecks. These latest results show that weak gravitational lensing is likely to play a very important role in this effort.

Max Tegmark is in the Department of Physics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

e-mail: max@physics.upenn.edu

1. Wittman, D. M., Tyson, J. A., Kirkman, D., Dell'Antonio, I. & Bernstein, G. *Nature* **405**, 143–148 (2000).
2. Van Waerbeke, L. *et al.* <http://xxx.lanl.gov/abs/astro-ph/0002500>; *Astron. Astrophys.* (in the press).
3. Bacon, D., Refregier, A. & Ellis, R. <http://xxx.lanl.gov/abs/astro-ph/0003008>
4. Kaiser, N., Wilson, G. & Luppino, G. A. <http://xxx.lanl.gov/abs/astro-ph/0003338>
5. De Bernardis, P. *et al.* *Nature* **404**, 955–959 (2000).

Plant pathology

The rise of the hybrid fungi

Clive Brasier

The evolutionary potential of a population is proportional to its level of genetic variation¹, which is mainly generated by sexual recombination between genetically different individuals. But many fungal plant pathogens, including the infamous potato blight, Dutch elm disease and chestnut blight pathogens, are clonal or part clonal outside their centres of origin — that is, their populations have low genetic diversity². Their evolutionary potential, therefore, is limited. Or is it?

A fast track to rapid evolution is hybridization between species, which combines the potential of two (or more) genomes. Until 1994, records of natural species hybrids in fungi were rare, with fewer than ten clear examples^{3,4}. However, writing in *Mycological Research*, Newcombe *et al.*⁵ describe the appearance of a hybrid rust fungus that parasitizes poplar trees. This is the latest in a series of reports of emerging fungal hybrids^{6–12}, all involving plant pathogens (Fig. 1).

Fungi from the same geographical area often exhibit strong genetic barriers (sexual or asexual) to interspecific hybridization. These barriers probably maintain gene combinations conferring optimal adaptation to the ecological niche¹³. They may also prevent

the spread of harmful genetic elements, such as viruses, between species. Such barriers might not exist, or might be weaker, between fungi that have remained geographically isolated. So — in theory — there is a greater possibility of hybridization when fungi spread beyond their normal geographical ranges.

Human influences have increased the chances of such occurrences. Immigration of fungal pathogens into new areas is likely to bring them into contact with resident, related species, especially those with a similar host or vector type. The potential for hybridization between residents and immigrants, however, will depend upon many other factors: the frequency of niche contact; the nature of any genetic barriers to hybridization; the degree to which their genomes can recombine; and the ability of any resulting hybrids to compete with the 'parent' species. Possible outcomes range from the acquisition of a single gene by one parent to a full species hybrid incorporating the genomes of both parents⁴.

With plant pathogens, the upshot could be transfer of pathogenic traits, or even the emergence of new plant diseases. This point is strikingly illustrated by Newcombe *et al.*⁵



Figure 2 Rust attack. Left, leaf from a poplar clone infected with *Melampsora medusae*. Right, leaf from a clone carrying a resistance gene showing flecks where *M. medusae* has failed to establish. But this clone is susceptible to attack by hybrids between *M. medusae* and *M. occidentalis*.

They report natural hybrids between the rust fungi *Melampsora medusae*, which occurs mainly on *Populus deltoides* in the eastern United States, and *M. occidentalis*, which occurs on *P. trichocarpa* in the west. In the 1980s, commercial poplar clones (*P. deltoides* crossed with *P. trichocarpa*), bred for resistance to *M. occidentalis*, were planted in the Pacific northwest; but in 1996, *M. medusae* × *M. occidentalis* rust hybrids appeared on these clones (Fig. 2), including some rust hybrids evidently back-crossed to *M. medusae*. The hybrids can attack *P. deltoides*, the commercial clones and even the local *P. trichocarpa*. So *M. medusae* pathogenicity genes could now be transferred into the resident *M. occidentalis* population⁵. Alternatively, the hybrids might supplant *M. occidentalis*.

The pathogenicity issue is further highlighted by new hybrids of the genus *Phytophthora* that are killing riverside alders in Europe. These are hybrids between *Phytophthora cambivora*, a common pathogen of hardwood trees, and a *Phytophthora* close to *P. fragariae*, a pathogen of strawberry and raspberry¹⁰. *Phytophthora* has not previously been recorded on alder, and tests show that neither *P. cambivora* nor *P. fragariae* is an aggressive pathogen of alder. In this case, therefore, the hybrids may have acquired the ability to exploit a new host — a possibility that had already been demonstrated with laboratory-generated *Phytophthora* hybrids¹⁴. The ecological significance of this process is evident, but how genes for novel host specificities are 'unmasked' or 'created' by hybridization is, as yet, unknown.

How hybrid genomes become stabilized and how 'fitness costs' influence hybrid survival is also poorly understood. In the *Phytophthora* and *Melampsora* examples^{5,10}, swarms of hybrids occur that are highly diverse both genetically and in their physical and behavioural characteristics (phenotypes). In *Melampsora*, study of the hybrid swarms has led to the discovery of an older hybrid population that, overall, displays an even greater adaptive range to its poplar

Fungal genus	Fungal group	Host and disease	Location	First report
<i>Melampsora</i>	Bm	Poplar leaf rust	New Zealand, South Africa ^{6,12}	1994
<i>Heterobasidion</i>	Bm	Conifer root rot	USA (California) ⁷	1996
<i>Ophiostoma</i>	Am	Dutch elm disease	Europe and southwest Asia ^{8,11,16}	1998
<i>Phytophthora</i>	Om	<i>Primula</i> , <i>Spathiphyllum</i> root rot	Netherlands ⁹	1998
<i>Phytophthora</i>	Om	Alder bark necrosis	Western Europe ¹⁰	1999
<i>Melampsora</i>	Bm	Poplar leaf rust	USA (Pacific northwest) ⁵	2000

Figure 1 Emerging hybrid fungal pathogens of plants. (Bm, basidiomycetes; Am, ascomycete; Om, oomycete fungi, now commonly assigned to Kingdom Chromista.)

hosts⁵ (see also ref. 12). In the alder phyto-phthoras, a common hybrid type predominates and is apparently still evolving, showing evidence of recombination and homogenization in its multiple-copy ribosomal DNA genes¹⁰. Other types are genomically unstable, showing remarkable, even continuous variation in their asexual lines. Many such variants are likely to be evolutionary dead ends, equivalent to Goldschmidt's 'hopeful monsters'¹⁵. Others, by chance, may succeed and become genetically refined under the influence of selection.

So are we experiencing a surge in pathogen evolution through interspecific hybridization, influenced mainly by human activity? Certainly we are in an era of increasing international movement of plants, and their associated pathogens; and one of increasing ecological disturbance and stress effects on plants, including habitat disturbance, climate change, pollution, geographically 'off-site' planting and crop breeding. Such 'episodic selection' conditions⁴ will increase the frequency of encounters between different pathogens. They may also present hybrids (even 'hopeful monsters') with greater opportunity for expression, by providing access to new or weakened hosts^{4,16}. In the case of *Melampsora* described by Newcombe *et al.*⁵, for example, traditional poplar breeding has facilitated hybrid survival.

Intriguingly, in five of the six new cases (Fig. 1) the species hybrids were initially

detected because of their unusual phenotypes, not through molecular genetic analysis (although such analysis was a critical confirmatory tool). This seems appropriate. Natural selection, too, acts on phenotypes. The evolutionary potential of a population is proportional to its level of phenotypic, as well as its genotypic, variation¹. It is now apparent that hybridization between fungal pathogens can produce a devastating array of phenotypes. ■

Clive Brasier is at the Forestry Commission Research Agency, Alice Holt Lodge, Farnham, Surrey GU10 4LH, UK.

e-mail: c.brasier@forestry.gov.uk

1. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon, Oxford, 1930).
2. Kohn, L. M. *Can. J. Bot.* **73**, S1231–S1240 (1995).
3. Burnett, J. H. *Trans. Br. Mycol. Soc.* **81**, 1–14 (1983).
4. Brasier, C. M. *Can. J. Bot.* **73**, S1213–S1221 (1995).
5. Newcombe, G., Stirling, B., McDonald, S. & Bradshaw, H. D. *Mycol. Res.* **104**, 261–274 (2000).
6. Spiers, A. G. & Hopcroft, D. H. *Mycol. Res.* **98**, 889–903 (1994).
7. Garbelotto, M., Ratcliff, A., Bruns, T. D., Cobb, F. W. & Orosina, W. J. *Phytopathology* **86**, 543–551 (1996).
8. Brasier, C. M., Kirk, S. A., Pipe, N. J. & Buck, K. W. *Mycol. Res.* **102**, 45–57 (1998).
9. Man in't Veldt, W. A. *et al.* *Phytopathology* **88**, 922–929 (1998).
10. Brasier, C. M., Cooke, D. E. L. & Duncan, J. M. *Proc. Natl Acad. Sci. USA* **96**, 5878–5883 (1999).
11. Abdelali, E., Brasier, C. M. & Bernier, L. *Mol. Plant–Microbe Interact.* **12**, 6–15 (1999).
12. Frey, P., Gatineau, M., Martin, F. & Pinon, J. *Proc. Int. Poplar Symp. III* 13–17 Sept. 1999, 34 (INRA, Orleans, 1999).
13. Dobzhansky, T. *Genetics and the Origin of Species* (Columbia Univ. Press, New York, 1937).
14. Ersek, T., English, J. T. & Schoelz, T. E. *Phytopathology* **85**, 1343–1347 (1995).
15. Goldschmidt, R. *Rév. Scient.* **86**, 607–612 (1948).
16. Brasier, C. M. *Bioscience* (in the press).

Statistical mechanics

Exploring phase space

Michael F. Shlesinger

The kinetic theory of gases illustrates how random collisions between hard spheres can lead to macroscopic properties such as pressure and temperature. The spheres perform simple random walks and travel a known 'average' distance between collisions. A paper by S. G. Cox and G. J. Ackland (*Physical Review Letters* **84**, 2362–2365; 2000) holds some surprises for the physics of colliding particles. By studying simple elastic collisions between three particles on a ring, the authors discover complex random-walk behaviour in a mathematical space (called a phase space) that tracks momentum exchanges. Although the phase space is a simple hexagonal lattice, the path and rate of exploration of this phase space varies greatly, depending on the mass ratios of the particles. This complexity of behaviour has implications for the question of how underlying microscopic behaviour can be averaged to obtain macroscopic laws. This question has a long history in statistical mechanics based around the concept of ergodicity.

Statistical mechanics measures an ensemble average, which means repeating an experiment many times and averaging the results. Physicists would rather study time series and have developed many methods for analysing them. For example, measuring the voltage as a function of time across a resistance can be studied through its power spectrum or wavelet transform. The mathematical condition allowing one to equate a time-series average with an ensemble average is called ergodicity. The idea is that, with time, a system will pass through almost all of its possible position and momentum states, allowing a time and ensemble average to be equivalent. The ergodic condition was first used in this context by Paul Ehrenfest in 1911 and mathematically refined by Garrett Birkhoff in 1931. Ehrenfest was seeking to explain, at a fundamental level, Ludwig Boltzmann's ideas on how microscopic behaviour underlies macroscopic laws. Ergodicity was never proven to be generally valid, although it remains a central assumption

of statistical mechanics. Most dynamical systems of practical interest are not ergodic.

One can always plot a particle's trajectory in three dimensions. This is informative, but more information is available if one plots the trajectory of a particle's momentum and position as a function of time. This is the phase-space plot. Because each particle has three space and three momentum coordinates, phase space becomes a $6N$ dimensional space if there are N particles. Think of a ball bouncing around a phase-space box. If the ball maintains its shape and visits almost all the points in the box then the system would be ergodic. If the ball loses its shape and spreads to cover the whole box then this system is ergodic and mixing. Systems that are both ergodic and mixing are the ones that physicists have used to investigate microscopic dynamics. It is easier to average over the bland homogeneous phase space of a mixing system than to deal with the quirky, irregular dynamics that underlie large fluctuations in more complex systems.

Physicists also like simple models where everything is out in the open — you know what you have put into the model and what comes out. Cox and Ackland use a similarly appealing and simple dynamical model to illustrate the intricacies of phase space. In their phase space there are three hard particles confined to a one-dimensional ring. The particles move with fixed velocity until they collide and exchange momentum. There are three types of momentum exchanges, A, B and C, involving different pairs of particles, which can be mapped onto an infinite hexagonal lattice — the momentum phase space. See Fig. 1 (overleaf) for an example of a hexagonal phase space.

In Cox and Ackland's model, the main variables are the masses of the three particles. For most mass ratios, a random sequence of A's, B's and C's ensues. Even though the system is entirely dynamical, a particle will diffuse through the hexagonal lattice with a motion characteristic of a simple random walk — that is, the distance travelled is proportional to the number of collisions raised to the power n , where $n = 0.5$. When two of the masses are nearly equal, patterns of momentum exchanges can result — such as ABABABA — that represent correlated flights with a faster exploration of the lattice phase space ($n > 0.5$). Large mass differences can create sequences that follow loops in phase space leading a particle back to its starting point. This produces a random walk with a slower exploration of phase space ($n < 0.5$). In this way, Cox and Ackland calculate a contour map of the n -values for a wide range of mass ratios. Finding the various outcomes by intuition would be much more difficult.

The clarity of this model shows that, even for a simple phase space, the manner and

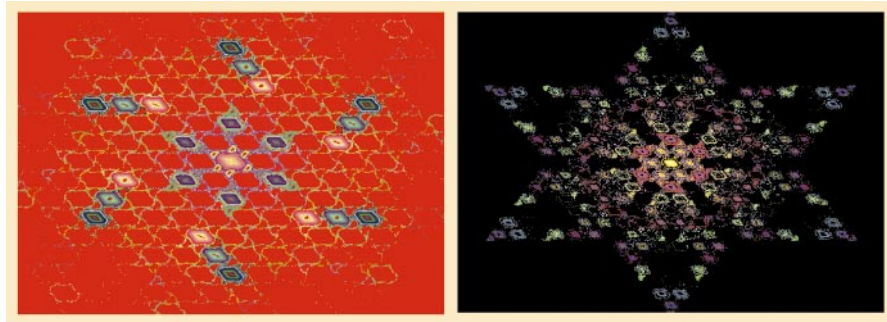


Figure 1 Exploring the phase space of a Zaslavsky map. The map represents a charged particle in a magnetic field having its stable closed orbits upset by kicks from an electromagnetic wave. A set of initial conditions was allowed to evolve and change colour with time, revealing complex and irregular flight patterns. Each image shows a different exploration of the same phase space. Most notable are trajectories moving on a hexagonal web allowing escape away from closed orbit regions. In their study, Cox and Ackland explore a hexagonal phase space corresponding to the motion of three particles on a ring. They find that the behaviour of such a system is not ergodic — that is, it cannot be assumed that the system always samples its phase space in the same way.

rate in which it is sampled can be fundamentally irregular and complex. What about more complicated phase spaces? We already know of dynamical systems that are not ergodic. Think of the Moon orbiting the Earth, while the Earth is orbiting the Sun. All the phase-space states of this three-body problem are not explored because chaotic behaviour is possible, as Henri Poincaré first recognized in 1892. What if the Earth–Moon–Sun system is perturbed? Could that make it ergodic? How would Newton explain the effect of his friend Halley's comet passing through this system? This problem was beyond Newton's grasp, and waited until the 1960s for Andrei

Kolmogorov, Vladimir Arnold and Jurgen Moser to show that phase space can simultaneously support stable closed orbits as well as a chaotic 'sea' of states. This finally provided an answer to the old problem of the stability of the Solar System. A related problem in modern physics is how to magnetically confine nuclei in a fusion reactor. Here, one wants to keep the particles confined to closed orbits and to prevent their escape into the chaotic sea.

Michael F. Shlesinger is in the Physical Sciences Division, Code 331, Office of Naval Research, 800 North Quincy Street, Arlington, Virginia 22217-5660, USA.
e-mail: shlesim@onr.navy.mil

Gene expression

Mutant weed breaks silence

Susan Tweedie and Adrian Bird

In many organisms, genomic DNA is decorated by the addition of methyl groups to certain cytosine residues. This phenomenon has proven hard to study in animals, in part because of the lack of genetically tractable model organisms. The models of choice — the fruitfly *Drosophila melanogaster* and the nematode worm *Caenorhabditis elegans* (to say nothing of 'surrogate animals' such as yeast) — lack all traces of methylated cytosine in their genomes. Fortunately this is not a problem in plants, where methylated genomes and powerful genetic methodologies coincide. The tiny weed *Arabidopsis thaliana* is the prime example, and researchers are making the most of this plant in their efforts to discover new facets of DNA methylation.

As in animals, DNA methylation in plants has been associated with genes that are silenced through their failure to be tran-

scribed into RNA. Genetic analysis has already identified one gene, *DDM1* (for 'decrease in DNA methylation'), that is somehow needed to maintain both methylation and silencing of genes^{1,2}. On page 203 of this issue³, however, Amedeo and colleagues show that silencing and methylation are not inextricably linked.

Foreign genes ('transgenes') can easily be introduced into plants, but often become silenced and acquire high levels of cytosine methylation of the regions that regulate gene expression (the promoters)⁴. To investigate this phenomenon, Mittelsten Scheid *et al.*⁵ had earlier mutated *Arabidopsis* plants carrying a silent, methylated transgene (called *HPT*) that encodes a protein conferring resistance to the drug hygromycin. They then looked for mutant plants that grew in the presence of hygromycin and had thus failed to silence the *HPT* transgene. They found

that all plants with reactivated *HPT* showed reduced methylation of this transgene, and that several of the mutations in such plants could be traced to the known *DDM1* gene². This genetic screen therefore provided a link between silencing and methylation, as mutation of a single gene (*DDM1*) affected both processes. But now, in a new screen, Amedeo *et al.*³ have identified a family of mutant plants in which the *HPT* transgene is reactivated while remaining heavily methylated. The authors have cloned the gene mutated in these plants and, to acknowledge its soporific effect on transcription, named it *MOM* (for 'Morpheus molecule'), after the Greek god of dreams.

Sequence analysis showed that *MOM* encodes a protein of 2,001 amino acids with some interesting relatives³. In particular, a region near the amino terminus shows some similarity to domains IV, V and VI of the helicase proteins of the SWI2/SNF2 family. These proteins regulate the access of transcription factors to genes in chromatin⁶ — the complex between chromosomal DNA and proteins. So the *MOM* protein may likewise be part of the chromatin-remodelling machinery in cells. This is a tantalizing possibility because the *DDM1* protein is also related to SWI2/SNF2 (ref. 2). Amedeo *et al.* speculate that *MOM* might encode half of the SWI2/SNF2 helicase domain, with the other half being supplied by a hypothetical binding partner. Whether or not this is so, the finding adds to a growing body of circumstantial evidence that silencing somehow involves chromatin remodelling.

The *MOM* protein could fit into the silencing mechanism in one of two ways. In the first scenario (Fig. 1a, overleaf), *MOM* is a mediator that acts downstream of the methylation signal to bring about silencing. Its role might be analogous to that proposed for the methyl-CpG-binding proteins of mammals⁷, which bind to methylated sites in DNA and recruit silencing proteins that can alter the structure of chromatin. In mammals, this is accomplished by the removal of acetyl groups from histones, the major protein component of chromatin.

In the second scenario (Fig. 1b), *MOM* induces silencing independently of methylation. *MOM* must act downstream of *DDM1*, as *ddm1* mutations both relieve silencing of the *HPT* transgene and affect DNA methylation⁵. A striking feature here is that methylation itself — although dependent on *DDM1* — would be irrelevant to transgene shutdown. Existing evidence favours this view, as methylation of *HPT* can be drastically reduced in plants that lack the methylating enzyme DNA methyltransferase, without concomitant reactivation of *HPT*⁵. Clearly, the mere presence of methylation at an inert plant gene cannot be taken to mean that methylation is the cause of silencing.

Several other findings hint at as-yet

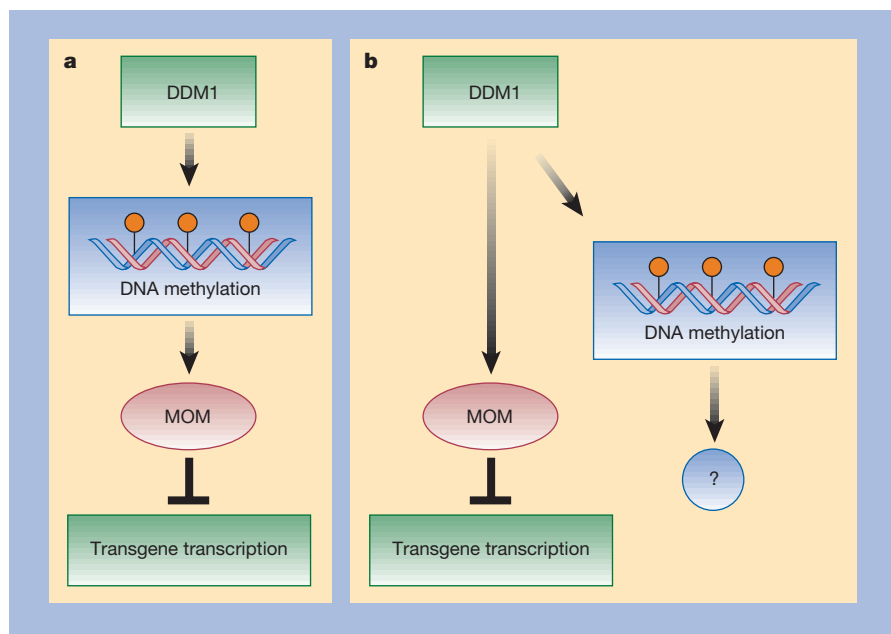


Figure 1 Two models for the role of the MOM protein in transgene silencing in *Arabidopsis*, as proposed by Amedeo *et al.*³. **a**, The DDM1 protein promotes DNA methylation (represented by circles linked to the DNA strand). Methylation exerts a negative effect on transcription through a mediator, MOM. **b**, The DDM1 protein has two downstream effects in distinct pathways: it promotes DNA methylation and it acts through MOM, in a DNA-methylation-independent manner, to repress transcription. In this scenario, methylation would not contribute to silencing of the transgene.

unfathomed complexity in plant gene-silencing mechanisms. Inactivation of a plant DNA methyltransferase gene (*MET1*) by 'antisense' RNA (which can bind to the 'sense' RNA produced from the gene itself) reduces global genomic methylation levels by up to 90% (refs 8, 9), but paradoxically leads to increased methylation and concomitant silencing of certain endogenous genes^{10,11}. In the same vein, it is surprising that none of the mutations that relieve silencing have so far been ascribed to loss of a DNA methyltransferase. The physical defects of the silencing mutants are also impressively mild. The *ddm1* mutants show abnormalities, but only after several rounds of self-pollination¹², whereas *mom* mutants have yet to show any physical defects after nine generations of self-pollination³.

Despite the breakthroughs, we are left with many more questions than answers. Do *mom* mutations reactivate silenced endogenous genes as well as the *HPT* transgene? Is the cauliflower mosaic virus 35S promoter — used in almost all such transgenic studies — really a good indicator of general mechanisms of gene silencing? Is there another half to the putative MOM DNA helicase, and can it really remodel chromatin? Do both plants and animals, as recently suggested¹³, use ATP-burning chromatin-remodelling engines to control gene silencing? Exactly what is the connection between DNA methylation and the SWI2/SNF2 proteins? The answers to some of these questions will doubtless come from more genetic screens in *Arabidopsis*. Let's hope that we do not need all 999 siblings of

Morpheus to demystify DNA methylation. ■ Susan Tweedie and Adrian Bird are at the Institute of Cell and Molecular Biology, University of Edinburgh, The Kings Buildings, Edinburgh EH9 3JR, UK.

e-mails: s.tweedie@ed.ac.uk

a.bird@ed.ac.uk

1. Vongs, A., Kakutani, T. J., Martienssen, R. A. & Richards, E. J. *Science* **260**, 1926–1928 (1993).
2. Jeddeloh, J. A., Stokes, T. L. & Richards, E. J. *Nature Genet.* **22**, 94–97 (1999).
3. Amedeo, P., Habu, Y., Afsar, K., Mittelsten Scheid, O. & Paszkowski, J. *Nature* **405**, 203–206 (2000).
4. Finnegan, E. J., Genger, R. K., Peacock, W. J. & Dennis, E. S. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49**, 223–247 (1998).
5. Mittelsten Scheid, O., Afsar, K. & Paszkowski, J. *Proc. Natl Acad. Sci. USA* **95**, 632–637 (1998).
6. Muchardt, C. & Yaniv, M. *J. Mol. Biol.* **293**, 187–198 (1999).
7. Bird, A. P. & Wolffe, A. P. *Cell* **99**, 451–454 (1999).
8. Finnegan, E. J., Peacock, W. J. & Dennis, E. S. *Proc. Natl Acad. Sci. USA* **93**, 8449–8454 (1996).
9. Ronemus, M. J., Galbiati, M., Ticknor, C., Chen, J. & Dellaporta, S. L. *Science* **273**, 654–657 (1996).
10. Jacobsen, S. E. & Meyerowitz, E. M. *Science* **277**, 1100–1103 (1997).
11. Jacobsen, S. E., Sakai, H., Finnegan, E. J., Cao, X. & Meyerowitz, E. M. *Curr. Biol.* **10**, 179–186 (2000).
12. Kakutani, T., Jeddeloh, J. A., Flowers, S. K., Munakata, K. & Richards, E. J. *Proc. Natl Acad. Sci. USA* **93**, 12406–12411 (1996).
13. Gibbons, R. J. *et al. Nature Genet.* **24**, 368–371 (2000).

Errata

In the article "Ornithology: British birds by number" (Robert M. May *Nature* **404**, 559–560; 2000), the examples of birds occurring through deliberate introduction and natural colonization were transposed in the *Nature* office. Little owls were a deliberate introduction, purple sandpipers a natural colonization.

Likewise, in "Palaeoanthropology: Coasting out of Africa" (Chris Stringer *Nature* **405**, 24–27; 2000), mention of a Figure 3 should not have appeared.

Daedalus

Moving pictures

Last week Daedalus devised a phosphorescent paint which absorbed and re-emitted light intermittently. When enough trapped unpaired electrons had accumulated in a crystal of its pigment, they switched a bistable magnetic inclusion which encouraged their radiative decay, and flipped back again when they had done so. The result was DREADCO's 'Wink' range of oscillating-brightness inks and paints.

Daedalus now realizes that their brightness varies spatially too. Even the most uniform Wink surface must vary in switching frequency from point to point, if only from local variations of temperature and ambient illumination. So its flash phasing soon ceases to be uniform; some regions will be bright when adjacent ones are dark. The boundaries between such regions will tend to drift, for the magnetic field at an edge will influence the strip of Wink just outside it. Daedalus is exploring the sort of flicker-shapes that can form, and how they change under different patterns of illumination.

The simplest instability is that of a long-period Wink surface on which regions of brightness form and wander, their shapes influenced by the changing illumination around them. As wallpaper or decor, such a surface would have the entrancing appeal of cloud-shadows moving over a landscape. As a kinetic camouflage or mutable dazzle-pattern, it might wonderfully obscure military vehicles — or indeed stolen cars — by distracting the observers' gaze.

But the most lucrative use would be in poster advertising. A cunning enough initial pattern, controlled by a grid of many programmed heating coils under the surface, would develop entrancingly. It could tell a picture story, or dissolve into random swirls from which slogans or images would emerge at intervals, or even reverse itself by a sort of spin-echo recapitulation. And by changing the heating program, the poster could be switched to a whole new story.

Wink patterns of higher frequency, in the flicker-fusion range, could induce new visual effects. A region winking at 50 Hz (say), on a ground winking at 51 Hz, should form a ghostly, subliminally compelling flicker-picture. Cunning flicker-field advertisements, all the more persuasive for their subtlety, should propel the consumer boom to new heights.

David Jones

The Further Inventions of Daedalus is published by Oxford University Press.

Lie detection and language comprehension

People who can't understand words are better at picking up lies about emotions.

People are usually no better than chance at detecting lies from a liar's demeanour^{1,2}, even when clues to deceit are evident from facial expression and tone of voice³. We suspected that people who are unable to understand words (aphasics) may be better at spotting liars, so we tested their performance as lie detectors. We found that aphasics were significantly better at detecting lies about emotion than people with no language impairment, suggesting that loss of language skills may be associated with a superior ability to detect the truth.

We studied the lie-catching abilities of ten patients who could understand individual words but who suffered severe deficits in comprehending spoken sentences after damage to the left cerebral hemisphere (LH). Their performance was compared with that of ten patients with damage to the right cerebral hemisphere (RH), ten healthy controls (C) and 48 undergraduates from the Massachusetts Institute of Technology (UC). Subjects watched a videotape in which each of ten people was shown twice consecutively: once attempting to conceal powerful negative emotions and once honestly revealing positive emotions. The sequence of the two interviews was random. Behavioural measurement showed that the interviews differed in subtle facial expressions and in pitch changes in the voice⁴.

Aphasics were significantly more accurate than controls at detecting lies. The mean of the LH group (0.61; s.d. = 0.10) was higher than that of the RH, C or UC groups (0.44, 0.47 and 0.46 respectively with standard deviations of 0.11, 0.16 and 0.16). An analysis of variance of the three matched groups (LH, RH and C) found a significant difference among groups ($F_{2,26} = 4.33$, $P < 0.03$). A planned contrast analysis comparing the LH group against the others was statistically significant ($F_{1,26} = 12.95$, $P < 0.002$) and the residual variance in the main effect was not significant; t -tests comparing the means of all four groups against the value of 5 (the value obtained by chance) revealed that only the LH group scored better than chance ($F = 9.00$, $P < 0.02$).

We then compared the performance of all groups on items where the clues were in facial expression (3 items), in pitch changes in the voice (1 item) or in the face and voice (6 items). LH patients do better when clues are in facial expression but not when the clues are from the voice alone (Table 1).

Our results support the untested claim that aphasic patients are unusually sensitive to deceitful behaviour^{5,6}. Perhaps damage to

Table 1 Success in interpreting lying cues

Group	Vocal pitch cues only	Facial expression cues only	Facial and vocal cues
LH	0.30	0.73	0.60
RH	0.20	0.50	0.45
C	0.20	0.57	0.47
UC	0.32	0.50	0.47

Values represent proportion correctly identifying liars. LH, left-hemisphere-damaged aphasics, mean age 58.4 years, patients at the Massachusetts General Hospital who gave informed consent. Their diagnoses, based on neurological examinations and MRI, were left middle cerebral artery infarct (nine patients) and subarachnoid haemorrhage (one subject). Neuropsychological testing revealed at least low average intellectual and perceptual abilities. Subjects achieved 95% correct (87.5–100% range) on a word-to-picture matching task and 89% correct on a lexical decision task (78–94% range)⁹, indicating recognition of single words. However, they performed at near-chance levels on a sentence-to-picture matching task, with an average accuracy of 58% (53–69% range)¹⁰, suggesting severely compromised comprehension of sentences. RH, right-hemisphere-damaged patients, mean age 59.6 years. C, matched controls, mean age 60.2 years. Both RH and C groups had equal numbers of men and women, were matched with the LH patients for education and IQ scores, were patients at the Massachusetts General Hospital, and had given informed consent. UC, undergraduate controls.

the circuitry underlying language comprehension results in the growth of compensatory skills in recognizing non-verbal behaviour. All but one of our aphasic patients were tested more than one year post-injury (the one who was tested within a year scored no better than chance). Although we cannot distinguish whether our patients were better lie detectors or simply better at detecting subtle cues to emotion, aphasics' abilities to recognize non-subtle facial expressions have, to our knowledge, never before been shown to be superior to those of controls^{7,8}. The superiority of aphasics to normal persons in any task is a rarity.

Nancy L. Etcoff*, Paul Ekman†, John J. Magee‡, Mark G. Frank§

*MGH East, Department of Psychiatry, Building 149, 13th Street, Charlestown, Massachusetts 02129, USA
e-mail: etcoff@mediaone.net

†University of California, Human Interaction Laboratory, 401 Parnassus Avenue, San Francisco, California 94143, USA

‡Circle.com, 725 Concord Avenue, Cambridge, Massachusetts 02138, USA

§Rutgers University, School of Communication, Information and Library Studies, 4 Huntington Street, New Brunswick, New Jersey 08901, USA

- Ekman, P. *Telling Lies* 2nd edn (Norton, New York, 1992).
- Ekman, P. & O'Sullivan, M. *Am. Psychol.* **46**, 913–920 (1991).
- Ekman, P., O'Sullivan, M., Friesen, W. V. & Scherer, K. R. J. *Nonverb. Behav.* **15**, 125–135 (1991).
- Ekman, P., Friesen, W. V., O'Sullivan, M. J. *Pers. Soc. Psychol.* **54**, 414–420 (1988).
- Head, H. *Aphasia and Kindred Disorders of Speech* (Cambridge Univ. Press, 1963).
- Sacks, O. *The Man Who Mistook His Wife for a Hat* (Summit, New York, 1985).
- Etcoff, N. L. in *Advances in Clinical Neuropsychology* (eds Goldstein, G. & Tarter, R. E.) 127–179 (Plenum, New York, 1986).
- Young, A. W. et al. *Brain* **116**, 941–959 (1993).
- Caplan, D. *Language* (MIT Press, Cambridge, MA, 1992).
- Rochon, E., Waters, G. S. & Caplan, D. *Brain Lang.* **46**, 329–349 (1992).

Growth factors

Formation of endothelial cell networks

The growth factor VEGF (vascular endothelial growth factor) promotes the formation of blood vessels in a process known as angiogenesis by inducing the proliferation and migration of endothelial cells¹. We show here that VEGF has another proangiogenic function — it can stimulate the elongation, network formation and branching of non-proliferating endothelial cells in culture that are deprived of oxygen and nutrients. As endothelial cells in tumours are exposed to chronic or intermittent hypoxic conditions^{2,3}, we propose that autocrine endothelial VEGF contributes to the formation of blood vessels in a tumour and promotes its survival.

Human umbilical-vein endothelial cells (HUVECs) and bovine adrenal cortex capil-

lary endothelial cells were cultured in a sandwich system⁴, in which the medium can only reach the cells from the edges of the culture (Fig. 1a). The combined processes of diffusion, consumption of oxygen and nutrients, and production of metabolites establish microenvironmental gradients across the width of the culture, like those that occur in tumours *in vivo*^{4–6}.

These gradients cause the cells to change shape and to reorganize themselves into networks (Fig. 1b,c). HUVECs at a distance of 0–2 mm from the edge of the sandwich cultures, where nutrients and oxygen are plentiful, retain their homogeneous, intact monolayer configuration (Fig. 1b). The monolayer morphology of HUVECs further in (3–5 mm from the edge) is disrupted and some endothelial networks are evident (data not shown). At the most hypoxic interior of the sandwich, 10–12 mm from the edge, these networks are fully formed (Fig. 1c). Control cultures, which had no top slide

and thus no gradients, kept their homogeneous monolayer configuration. Expression of VEGF protein increased starting from the edge of the sandwich culture and peaked in the central O_2 /nutrient-poor region (Fig.

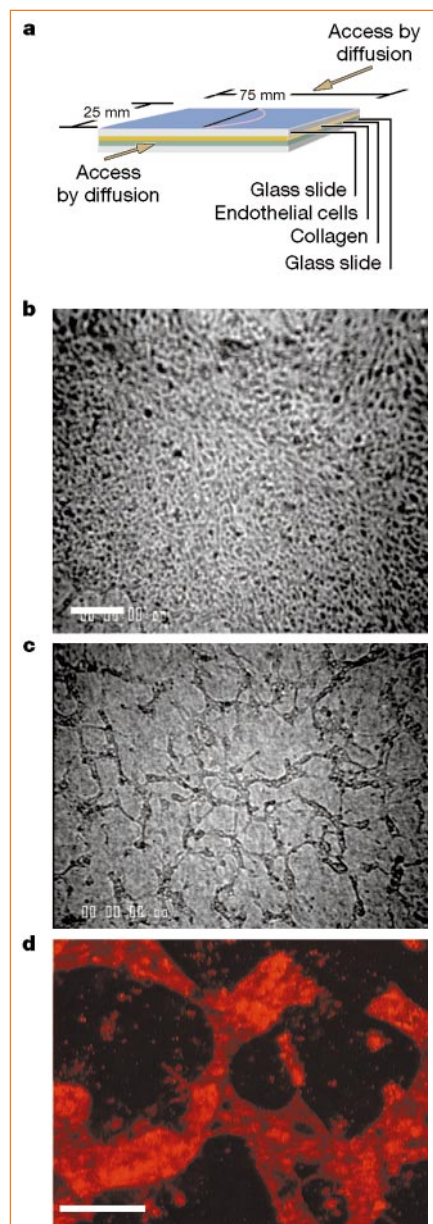


Figure 1 VEGF produced by endothelial cells (ECs) promotes network formation in the hypoxic region of tumour-mimetic sandwich cultures. **a**, Sandwich culture: ECs are plated onto a collagen-coated slide (75 × 25 mm) at the bottom; a gap of $400 \pm 40 \mu\text{m}$ separates the top and bottom slides, permitting access of oxygen and nutrients only from the edges of the sandwich⁴. The black line across the culture represents the scanning line for measurements. **b,c**, HUVEC morphology as a function of position within the sandwich culture 24 hours after installing the upper slide; scale bar, $400 \mu\text{m}$; view at 0–2 mm from the edge (**b**); 10–12 mm in from the edge (centre of the culture) (**c**). **d**, Immunohistochemical staining of VEGF in HUVECs, 11 mm from edge; scale bar, $40 \mu\text{m}$. Cultures were fixed in cold acetone and methanol. VEGF was detected by a mouse anti-human VEGF monoclonal antibody (PharMingen) and a Cy3-conjugated secondary goat anti-mouse IgG antibody (Jackson). Similar results were obtained for BCE cultures (see Supplementary Information).

1d), an effect also seen for stromal cells in sandwich culture⁶.

To investigate the temporal and spatial dynamics of this process, we measured the partial pressure of oxygen (pO_2), the concentration of VEGF, and the segment and branch lengths of developing networks after 0, 1, 1.5, 3, 6, 9 and 24 hours in sandwich culture (Fig. 2, and Supplementary Information). Pronounced gradients of pO_2 were created after 1 hour's culture, with cells on the interior experiencing oxygen levels below 30 mm Hg, dropping to about 5 mm Hg after 1.5 h (Fig. 2a).

The oxygen gradient induced a gradient of VEGF expression in the opposite direction (Fig. 2b). By 1.5 h, there was only a moderate increase in VEGF expression apparent in the interior, with no evidence of endothelial networks. VEGF gradients were clearly established at 3 h, while networks were only partially formed; networks then progressed to full formation over the next 6 h under minimal pO_2 (Fig. 2c,d). Networks and pO_2 gradients were produced even in the absence of glucose or serum gradients (see Supplementary Information). Our results indicate that hypoxia-induced endothelial VEGF drives these cells to form into networks in the absence of other cell types or exogenous growth factors.

To determine whether network formation was directly mediated by endothelium-produced VEGF, we added an anti-VEGF

neutralizing antibody (A4.6.1, Genentech; $19.3 \mu\text{g ml}^{-1}$) to sandwich cultures before positioning the upper slide. After 9–10 h, networks were fully developed in untreated HUVEC sandwich cultures, but, despite the gradient in oxygen tension, were absent in cultures that had undergone anti-VEGF treatment (see Supplementary Information). VEGF thus appears to be necessary for the formation of endothelial networks under oxygen/nutrient-gradient conditions.

VEGF acts as a mitogen on endothelial cells⁷ and as a survival-promoting factor on immature blood vessels^{1,7}. We have shown that it can also promote endothelial network formation without requiring cell proliferation, by inducing the elongation and reorganization of endothelial cells. The short time required for network formation (6–7 h) would not allow significant proliferation to occur (Fig. 2). Also, cultured HUVECs grow maximally at 7.5–10% O_2 (pO_2 , 57–76 mm Hg)⁸, tensions that exist only at the edge of the sandwich culture, where no networks form: networks appeared instead in the most hypoxic regions where cells do not proliferate⁴.

Vascular sprouting and remodelling can occur *in vivo* in both normal⁹ and tumour¹⁰ tissues without significant proliferation of endothelial cells. Because these cells in tumours are exposed to intermittent and chronic hypoxic conditions *in vivo*^{2,3} and as they can upregulate VEGF under low oxy-

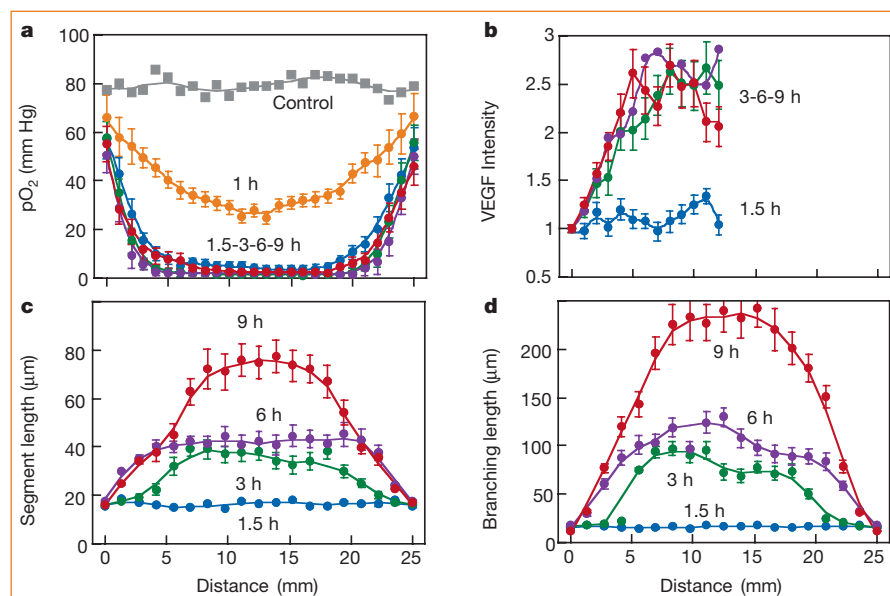


Figure 2 Measurement of spatial and temporal gradients of pO_2 and VEGF, and segment and branch lengths across the width of HUVEC sandwich cultures. Profiles are shown for pO_2 (**a**); intensity of VEGF staining (**b**); segment length (**c**); and branching length at 1 h (orange circles) (**d**), 1.5 h (blue), 3 h (green), 6 h (purple) and 9 h (red) after adding the upper slide; controls having no upper slide (grey squares); mean $\pm$ s.d. is shown ($n=3$). The sandwich construct was transparent⁴, permitting the measurement of pO_2 , VEGF expression and the degree of endothelial reorganization across the culture. Oxygen gradients were measured by using a high-resolution phosphorescence quenching microscopy technique^{2,13}. VEGF was detected by immunohistochemistry (Fig. 1) and mapped spatially by densitometry of digitized fluorescence images. VEGF profiles were normalized for the number of ECs, as estimated by CD31 staining (mouse anti-human CD31 monoclonal antibody; PharMingen). Endothelial network formation was assessed as a function of oxygen tension by simultaneously quantifying cell elongation and cell branching by transillumination and digital image analysis. A network segment was defined as a 'straight' line with no significant angular deviation ($< 8^\circ$). The cumulative branching length was the sum total of locally connected segments.

gen tensions *in vitro*^{11,12}, we propose that the hypoxia-driven autocrine stimulation of endothelial cells by VEGF in solid tumours helps promote the formation and reorganization of their vascular network. Other molecules present in the tissue microenvironment, such as nitric oxide and carbon monoxide¹², may also affect endothelial VEGF production and angiogenesis *in vivo*.
Gabriel Helmlinger*, **Mitsuhiro Endo***, **Napoleone Ferrara†**, **Lynn Hlatky‡**, **Rakesh K. Jain***

*Steele Laboratory, Department of Radiation Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA

†Genentech, South San Francisco, California 94080, USA

‡Department of Radiation Oncology, Dana-Farber Cancer Institute/Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02215, USA

1. Carmeliet, P. *Nature Med.* **6**, 389–395 (2000).
2. Helmlinger, G., Yuan, F., Dellian, M. & Jain, R. K. *Nature Med.* **3**, 177–182 (1997).
3. Kimura, H. *et al.* *Cancer Res.* **56**, 5522–5528 (1996).
4. Hlatky, L. & Alpen, E. L. *Cell Tissue Kinet.* **18**, 597–611 (1985).
5. Hlatky, L., Hahnfeldt, P., Tsiou, C. & Coleman, C. N. *Br. J. Cancer*, **74**, S151–S156 (1996).
6. Hlatky, L. *et al.* *Cancer Res.* **54**, 6083–6086 (1994).
7. Ferrara, N. & Davis-Smyth, T. *Endocr. Rev.* **18**, 4–25 (1997).
8. Nomura, M. *et al.* *J. Biol. Chem.* **270**, 28316–28324 (1995).
9. Sholley, M. M. *et al.* *Lab Invest.* **51**, 624–634 (1984).
10. Patan, S., Munn, L. & Jain, R. K. *Microvasc. Res.* **51**, 260–272 (1996).
11. Namiki, A. *et al.* *J. Biol. Chem.* **270**, 31189–31195 (1995).
12. Liu, Y. *et al.* *J. Biol. Chem.* **273**, 15257–15262 (1998).
13. Torres Filho, I. P. & Intaglietta, M. *Am. J. Physiol.* **265**, H1434–H1438 (1993).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Pollution

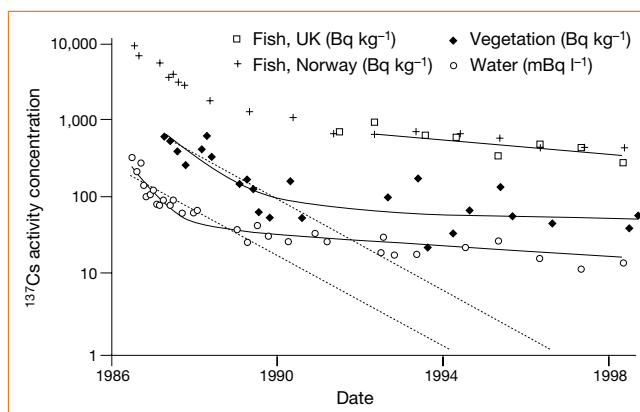
Chernobyl's legacy in food and water

Radiocaesium (¹³⁷Cs) from the 1986 Chernobyl accident has persisted in freshwater fish in a Scandinavian lake for much longer than was expected¹. On the basis of new data generalizing this observation, we propose that the continuing mobility of ¹³⁷Cs in the environment is due to the so-called 'fixation' process of radiocaesium in the soil tending towards a reversible steady state. Our results enable the contamination of foodstuffs by Chernobyl fallout to be predicted over the coming decades. Restrictions in the United Kingdom, for example, may need to be retained for a further 10–15 years — more than 100 times longer than originally estimated.

We have measured ¹³⁷Cs activity concentrations (Q_i) in terrestrial vegetation (seven sites, including data from ref. 2), lake water (dissolved phase, two lakes) and mature fish (three species) in Cumbria, UK, over the

Figure 1 Long-term changes in ¹³⁷Cs in brown trout in Norway (from ref. 1), and in perch, terrestrial vegetation and water in Cumbria, UK. The decline in ¹³⁷Cs in immature fish, water and vegetation during the first five years has an effective ecological half-life (T_{eff}) of 1–4 years^{1,4} as a result of 'fixation'. Dotted lines indicate the hypothetical continuation of irreversible fixation. Fits of the two-exponential model to our data, indicating the reversibility of 'fixation', are shown as solid lines.

Long-term declines in all three Cumbrian systems are similar (T_{eff} range of 6–30 years) and are in quantitative agreement with results from the Norwegian fish study¹.



same period as the Norwegian study¹. Our results for vegetation and water contamination (examples in Fig. 1) show the same two-component exponential decline ($Q_i = Q_1 e^{-k_1 t} + Q_2 e^{-k_2 t}$) observed for immature fish¹. The decline in ¹³⁷Cs in mature fish was influenced by slower biological uptake rates during the initial period after Chernobyl^{1,3}, so only the second component of the decline is shown for our fish data (Fig. 1).

Our results show that the effective ecological half-life (T_{eff} , the time for the ¹³⁷Cs concentration to reduce by 50%) in young fish, water and terrestrial vegetation has increased from 1–4 years during the first five years after Chernobyl^{1,4} to 6–30 years in recent years. The common rate of decline in ¹³⁷Cs concentration in lake water, fish and vegetation suggests that it is controlled by the same process in all three pools. This is consistent with a controlling influence of changes in chemical availability of ¹³⁷Cs in soil (in these lakes, long-term ¹³⁷Cs in the water originates in catchment runoff⁵).

The decline in ¹³⁷Cs mobility and bioavailability over the first few years after fallout is believed to be controlled by slow diffusion of ¹³⁷Cs into the illitic clay mineral lattice⁴. This 'fixation' process controls the amount of radiocaesium in soil water and therefore its availability to terrestrial biota and for transfer to rivers and lakes⁴. Studies of ¹³⁷Cs in contaminated sediments^{6,7}, however, indicate that this process may be reversible. From the persisting mobility of radiocaesium, and particularly the increase of T_{eff} towards the physical decay rate of ¹³⁷Cs ($T_{1/2} = 30.2$ years), we conclude that the sorption-desorption process of radiocaesium in soils and sediments is tending towards a reversible steady state.

The continuing mobility of ¹³⁷Cs in the environment means that foodstuffs will remain contaminated for much longer than was first expected. In the United Kingdom, restrictions on the sale and slaughter of sheep are currently in place on 389 upland farms (with about 232,000 sheep) on which

some sheep have ¹³⁷Cs activity concentrations above the UK limit for the entry of meat into the food-chain (1,000 Bq kg⁻¹). During our studies on three restricted farms in 1991–93 (ref. 8), the maximum ¹³⁷Cs level in sheep meat was 1,870 Bq kg⁻¹.

Assuming that this is typical of restricted farms within the UK and using the rates of long-term decline we have estimated, restrictions may need to remain in place on some farms for a total of 30 years after the Chernobyl accident, which is more than 100 times longer than initially expected. In some areas of the former Soviet Union, consumption of forest berries, fungi⁹ and fish¹⁰ (present ¹³⁷Cs content, 10–100 kBq kg⁻¹), which contribute significantly to people's radiation exposure, will need to be restricted for at least a further 50 years.

J. T. Smith*, **R. N. J. Comans†**, **N. A. Beresford‡**, **S. M. Wright‡**, **B. J. Howard‡**, **W. C. Camplin§**

*Centre for Ecology and Hydrology, Winfrith Technology Centre, Dorchester DT2 8ZD, UK
e-mail: jts@ceh.ac.uk

†Netherlands Energy Research Foundation (ECN), PO Box 1, 1755 ZG Petten, The Netherlands

‡Centre for Ecology and Hydrology, Grange-over-Sands LA11 6JU, UK

§Centre for Environment, Fisheries and Aquaculture Science, Pakefield Road, Lowestoft NR33 0HT, UK

1. Jonsson, B., Forseth, T. & Ugedal, O. *Nature* **400**, 417 (1999).
2. Watterson, J., Nicholson, K., Sandalls, J. & Pomeroy, I. *J. Environ. Radioactivity* (in the press).
3. Elliott, J. M., Hilton, J., Rigg, E., Tullett, P. A., Swift, D. J. & Leonard, D. R. *P. J. Appl. Ecol.* **29**, 108–119 (1992).
4. Smith, J. T. *et al.* *Environ. Sci. Technol.* **33**, 49–54 (1999).
5. Smith, J. T., Leonard, D. R. P., Hilton, J. & Appleby, P. G. *Health Phys.* **72**, 880–892 (1997).
6. Smith, J. T. & Comans, R. N. J. *Geochim. Cosmochim. Acta* **60**, 995–1004 (1996).
7. Comans, R. N. J. in *Mineral-Water Interface Reactions* 179–201 (ACS Symp. Ser. 715, Am. Chem. Soc., Washington DC, 1999).
8. Beresford, N. A., Barnett, C. L., Crout, N. M. J. & Morris, C. *Sci. Tot. Environ.* **177**, 85–96 (1996).
9. Beresford, N. A. & Wright, S. M. (eds) *Self-Help Countermeasure Strategies for Populations Living Within Contaminated Areas of the Former Soviet Union and an Assessment of Land Currently Removed from Agricultural Usage* 1–82 (Institute of Terrestrial Ecology, Grange-over-Sands, 1999).
10. Smith, J. T., Kudelsky, A. V., Ryabov, I. N. & Hadderingh, R. H. *J. Environ. Radioactivity* **48**, 359–369 (2000).

Detection of weak gravitational lensing distortions of distant galaxies by cosmic dark matter at large scales

David M. Wittman*, J. Anthony Tyson, David Kirkman, Ian Dell'Antonio† & Gary Bernstein‡

* Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey, 07574, USA

† Kitt Peak National Observatory, NOAO, Tucson, Arizona 85726, USA

‡ Astronomy Department, University of Michigan, Ann Arbor, Michigan 48109, USA

Most of the matter in the Universe is not luminous, and can be observed only through its gravitational influence on the appearance of luminous matter. Weak gravitational lensing is a technique that uses the distortions of the images of distant galaxies as a tracer of dark matter: such distortions are induced as the light passes through large-scale distributions of dark matter in the foreground. The patterns of the induced distortions reflect the density of mass along the line of sight and its distribution, and the resulting 'cosmic shear' can be used to distinguish between alternative cosmologies. But previous attempts to measure this effect have been inconclusive. Here we report the detection of cosmic shear on angular scales of up to half a degree using 145,000 galaxies and along three separate lines of sight. We find that the dark matter is distributed in a manner consistent with either an open universe, or a flat universe that is dominated by a cosmological constant. Our results are inconsistent with the standard cold-dark-matter model.

The large-scale distribution of dark matter depends both upon the nature of the dark matter and the global cosmological parameters that describe the Universe. Information on the large-scale distribution of matter is thus one of the primary goals of modern observational astronomy. To date, most of what we know about the large-scale structure of the Universe comes from the observed anisotropies in the cosmic microwave background (CMB) and from the distribution of galaxies. The CMB provides the earliest sample of mass fluctuations, from a time when the Universe was 100,000 times younger¹. Different cosmological models predict different scenarios in the growth of mass structures over cosmic time, so comparison of the CMB-derived mass spectrum with that seen at later times is a powerful test of cosmological models. The large-scale mass distribution at later times has traditionally been characterized through the large-scale distribution of galaxies on the assumption that light traces mass.

The distribution of this dark mass can be investigated more directly via its gravitational effects on the appearance of background galaxies. Any foreground mass bends light rays from a distant source, moving the apparent position of the source to a new position on the sky and stretching its image tangentially, by an amount proportional to the foreground mass. This weak lensing effect has already been used to study the mass distribution within clusters of galaxies, where the large mass associated with the clusters makes the gravitationally induced ellipticity of the background galaxies easily detectable^{2–7}. In principle, weak lensing can also tell us about large-scale structure through the cumulative effect of many intervening overdensities. A deep image of a patch of the sky looks out through the three-dimensional arrangement of galaxies seen in projection: any two galaxies are not likely to be physical neighbours and in the absence of lensing, their projected shapes or ellipticities are statistically uncorrelated. In the presence of foreground mass overdensities—that is, those occurring between distant galaxies and the observer—the light rays from galaxies narrowly separated on the sky travel similar paths past intervening mass concentrations and thus undergo similar image distortions. The resulting correlation of distant galaxy ellipticities is highest at small angular separation and drops for widely separated galaxies whose light bundles travel through completely different structures (Fig. 1). Different

cosmological models predict different behaviour for correlations of galaxy ellipticities versus angular separation on the sky.

Theoretical expectations for this cosmic shear on angular scales of 10–30 arcmin range from a few per cent for the standard cold-dark-matter model to less than one per cent for an open universe which would expand forever^{8–15}. The typical background galaxy has an

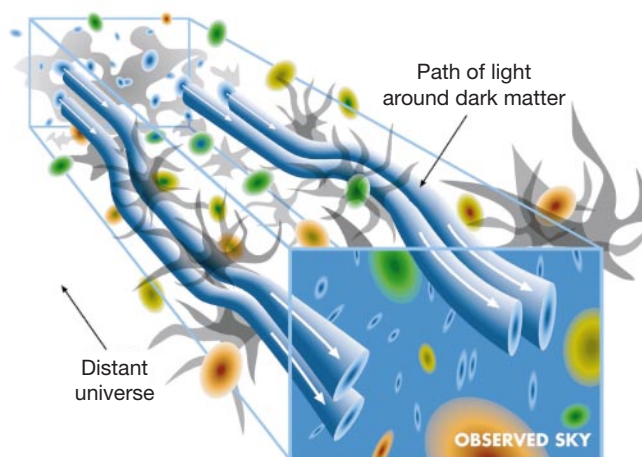


Figure 1 The distorted universe. Light rays from distant galaxies travel a tortuous path through a universe filled with clustering dark mass. Every bend in the path of a bundle of light from a distant galaxy stretches its apparent image. The orientation of the resulting elliptical images of galaxies contains information on the size and mass of the gravitational lenses distributed over the light path. The figure shows a schematic view of weak gravitational lensing by large-scale mass structure: distant galaxy orientation is correlated on scales characteristic of the lensing dark matter structures. Light bundles from two distant galaxies which are projected closely together on the sky follow similar paths and undergo similar gravitational deflections by intervening dark matter concentrations. Apparent orientations of distant galaxies are thus correlated on angular scales of less than a few degrees. The larger the mass in the gravitational deflectors, the larger the faint galaxy ellipticity correlations on a given angular scale. These ellipticity correlations of distant galaxies reveal the statistics of the large-scale dark matter distribution in the intervening universe—a central diagnostic of the underlying cosmology.

intrinsic ellipticity of roughly 30%, so that many thousands of source galaxies must be averaged together to detect the small change induced by cosmic shear. In addition, a large area of sky must be covered, because mass structures should span a few arcminutes to a degree at a typical mid-path distance of redshift 0.4 (about three billion light years). Earlier attempts to measure cosmic shear were inconclusive^{16–19}, with the main difficulty being control of systematic errors in galaxy shapes arising from the optical system or the process of observation. The earliest attempts with photographic plates, while covering a large field, suffered from plate-to-plate systematics as well as nonlinearity and lack of sensitivity. The sensitivity, linearity and reproducibility problems were solved with the advent of charge-coupled devices (CCDs), but the small field size covered by early CCDs was a problem. Mosaics of large CCDs now approach the desired one-degree field size, and are stimulating much activity in weak gravitational lensing.

We have imaged large areas of sky in several directions using a mosaic of CCDs on a large telescope, covering hundreds of thousands of distant galaxies at multiple wavelengths. We describe the steps taken to minimize systematic errors and to select 145,000 of the most reliable distant galaxy measurements. We find significant ellipticity correlations on angular scales of 0.04°–0.5°. This is, to our knowledge, the first direct probe of the aggregate mass distribution in the Universe at late times on the scale of several billion light years, and the results are consistent with two leading cosmological models.

Wide-field imaging with control of systematic shape errors

We observed three blank (that is, not containing any known mass concentrations) fields—at 23 h 48 min, +00° 57' J2000; 04 h 29 min, –36° 18'; and 11 h 38 min, –12° 33'—over a period of several years.

This was done using the Big Throughout Camera²⁰, which comprises an array of four large, blue-sensitive CCDs at the Cerro Tololo Inter-American Observatory's 4-m Blanco telescope. Constructed originally for weak lens observations, this camera covers a 35-arcmin field of view with 0.43-arcsec pixels. We took multiple 500-s exposures shifted by 5–7 arcmin and combined them to cover a 43-arcmin field. Before combining, we took several steps to reduce systematic errors arising from the optical system. First, we registered all the images onto a common linear coordinate system, free of the known radial distortion of the telescope optics. We then used the shapes of stars, which are foreground point sources free of the gravitational lensing effect, to correct any additional anisotropies in the point-spread function (the response of the optical system to point sources), such as those due to astigmatism and guiding errors.

The shape of a star or galaxy can be described by its second central moments, $I_{xx} \equiv \sum I w x^2$, $I_{yy} \equiv \sum I w y^2$ and $I_{xy} \equiv \sum I w xy$, where $I(x,y)$ is the intensity distribution above the night sky level, $w(x,y)$ is a weight function, the sum is over a contiguous set of pixels defined as belonging to the galaxy, and the coordinate system has been translated so that the first moments vanish. The second moments can be combined to form a size, $I_{xx} + I_{yy}$, and two components of a pseudo-vector ellipticity, $e_1 \equiv (I_{xx} - I_{yy})/(I_{xx} + I_{yy})$ and $e_2 \equiv 2I_{xy}/(I_{xx} + I_{yy})$, which vary in the range [–1,1] (ellipticity in its colloquial sense is the amplitude of this pseudo-vector, $\epsilon \equiv (e_1^2 + e_2^2)^{1/2}$ with its range [0,1]). Traditional intensity-weighted moments are calculated with $w = 1$, but this produces ellipticity measurements with noise properties that are far from optimal—or even divergent. In cases of white noise, the formal optimal weight for an elliptical source is a noise-free image of that elliptical source²¹.

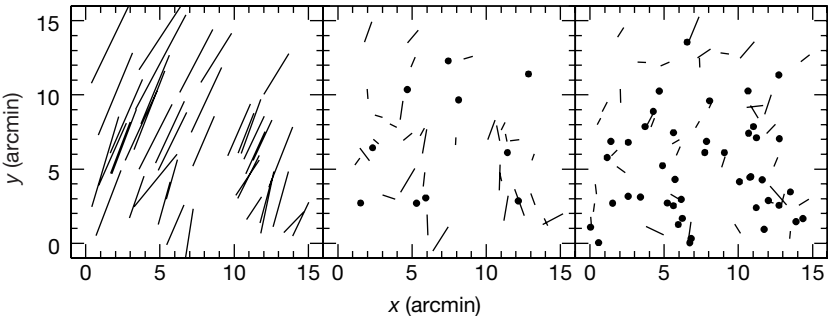


Figure 2 Making stars round. Foreground stars at many positions in the field of view are used to correct for position-dependent systematic ellipticity error. Convolution with a position-dependent kernel with ellipticity components equal and opposite to those of the stars reduces this systematic error everywhere in the field. Here we illustrate this technique with one particularly bad frame of raw data from one of the four CCDs in our mosaic. Each panel represents stars at their positions in the field as a line encoding the ellipticity and position angle, or as a point if the ellipticity is less than one per cent. The left panel is the raw data; the stars in a more typical frame have only half the ellipticity of those shown here, or about 5%. The middle panel shows the stellar shapes in the same single

image after convolution with the rounding kernel. The stars are vastly less out of round, but local correlations still exist. The right panel shows the final shapes of stars in the same region of sky, after combining ten shifted exposures, convolving the combined image, and measuring the shapes in more than one filter. Many of the local correlations in the middle panel have disappeared. The density of stars is greater due to better identification of stars in the combined image. The final field size is roughly ten times larger in area than this patch, and contains about 1,000 such stars in most of our mid-latitude fields. This figure is for illustration purposes only; Fig. 4 contains a quantitative assessment of the final level of systematic error.

Table 1 Summary of cosmological models

Model*	Ω_b	$\Omega_{\text{matter}} - \Omega_b$	Ω_Λ	H_0	n	σ_8	Normalization
Standard cold dark matter (blue)	0.05	0.95	0	50	1.0	1.17	COBE only
Λ -dominated, flat (green)	0.039	0.291	0.67	70	0.94	0.84	COBE + clusters
Open universe (orange)	0.045	0.405	0	65	1.01	0.71	COBE + clusters

These cosmological models were chosen in order to put our ellipticity correlation measurements in context. The old standard cold-dark-matter model, in which the Universe is nearly closed by cold dark matter, is also disfavoured in other observations. Its r.m.s. mass contrast is normalized to that found 300,000 years after the Big Bang via the cosmic microwave background radiation fluctuations observed with the Cosmic Background Explorer satellite (COBE). The other two models agree with a wide variety of observations, but only the Λ -dominated, (flat) model also agrees with the recent evidence from supernova studies for accelerated expansion. Ω_b is the fraction of critical density in ordinary (baryonic) matter, Ω_{matter} is the fraction in all matter (mostly dark matter), and Ω_Λ is the fraction in dark energy (the cosmological constant). H_0 is the Hubble constant in units of $\text{km s}^{-1} \text{Mpc}^{-1}$. The power spectrum $P(k)$ of mass density fluctuations is often plotted in terms of inverse size: the wavenumber k is inversely proportional to length. The parameter n is the slope of the primordial density power spectrum as a function of k : $P(k) \propto k^n$. For a scale-free power spectrum of density fluctuations, $n = 1$. The parameter σ_8 is the current r.m.s. mass contrast in a random sphere of radius $8(100/H_0) \text{ Mpc}$ compared with that for numbers of galaxies²⁶. The choice of $n = 1$ and COBE normalization for standard cold dark matter results in too much mass fluctuation on galaxy cluster scales. By adjusting the slope n and current r.m.s. mass contrast σ_8 , models can be forced to fit the r.m.s. mass contrast now on galaxy cluster scales as well as the COBE normalization.

* Colour in parentheses refers to the corresponding line in Fig. 5.

In the absence of such an image, weak lensing measurements are generally made with circular gaussian weights. We use an elliptical gaussian as the weight function, which places more weight on the inner parts of the galaxy image where the signal-to-noise ratio is higher, and is nearly optimal for most point-spread functions and for typical exponential galaxy profiles.

The moments of the gaussian weight ellipse are iterated (from initial values provided by unweighted moments) to match the size and shape of the object, in order to obtain the highest possible signal-to-noise ratio and to ensure that the measured ellipticity is not biased toward the shape of the weight function. This technique of adaptively weighted moments has been extensively tested on simulated and real data, and has been shown to be unbiased. On simulated data, the algorithm recovers a somewhat higher fraction of the artificially induced shear than does simple intensity weighting. However, our final results do not depend on this particular weighting scheme. Its real benefit lies in rejection of peculiar objects, the vast majority of which are overlapping galaxies seen in projection. If the final centroid of a galaxy differs significantly from the starting centroid, that galaxy is rejected. If the centroid does not shift, the galaxy is accepted (and probably suffers little contamination by its neighbour). Object candidates are also rejected if: the centroid or the ellipticity fails to converge; they are too near the edge of the image; the size grows too large; or the moments are negative. About one-third of candidates found by the detection software (which provides the initial unweighted moments and can ‘detect’ occasional noise peaks) are rejected. For candidates which survive, the measurement error in the final ellipticity is accurately estimated by propagating the poissonian photon noise through the moment equations.

We used foreground stars at many positions in the field of view to measure and correct for systematic ellipticity error. Stars are distinguished from galaxies by their clear separation at the bright

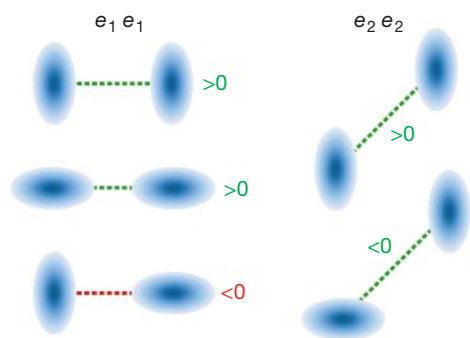


Figure 3 Lens-induced galaxy orientation correlations. Pairs of background galaxies, separated on the sky by some angle, can have their relative orientations affected by weak lensing. The ellipticity components of any galaxy i with respect to galaxy j can be visualized as $e_1 = \epsilon \cos(2\theta)$ and $e_2 = \epsilon \sin(2\theta)$, where ϵ is the scalar ellipticity and θ is the position angle with respect to a line joining the two galaxies. Ellipticity correlation functions are computed from the products of the ellipticity components of millions of such pairs, as a function of angular separation between pairs. A variety of relative orientations are illustrated along with their contributions to the correlations. Gravitational lensing leaves its signature on these correlations in several ways. First, the amplitude of the correlations scales with the amount of foreground mass. Second, the correlations are large at small separations and drop to nearly zero at large separations in a particular way. The bottom (red) case on the left cannot be caused by gravitational lensing, so that $\langle e_1 e_1 \rangle$ (averaged over many pairs) is always positive in the absence of systematic error. But lensing can cause the e_2 product to have either sign, as shown, and $\langle e_2 e_2 \rangle$ should become negative at a separation characteristic of the underlying cosmology. Third, correlations between e_1 and e_2 (not shown here) are not induced by gravitational lensing, so that any putative measurement of a weak lensing effect should vanish in the cross-correlation $\langle e_1 e_2 \rangle + \langle e_2 e_1 \rangle$.

end of a size/flux-density diagram. We identified roughly 100 such stars on each exposure of each CCD and made a least-squares fit (with 3σ cut-off) of a second-order polynomial to the spatial variation of their ellipticity components $e_1^*(x,y)$ and $e_2^*(x,y)$, which would be zero at all points in an ideal observation free of point-spread function anisotropy. Fischer and Tyson²² have shown that non-zero e_1^* and e_2^* can be cancelled by convolution with a small (three pixel square) flux-conserving kernel with ellipticity components equal and opposite to those of the stars. Simulations as well as weak lensing data on clusters of galaxies show that systematics induced in faint galaxies are also removed in this process of circularizing stars. We convolved each image with its resulting position-dependent circularizing kernel, after which the stellar ellipticities show little variation as a function of position. We then combined the images by averaging with a 3σ cut-off, and repeated the point-spread function rounding on the combined image (using roughly 1,000 stars and a fourth-order polynomial in this case). Figure 2 depicts the evolution of one of our worst raw images through this process.

Catalogues of distant galaxies

We repeated the observing and image processing for each field in three wavelength bands centred on 450 nm, 650 nm and 850 nm (ref. 23), and for two of the fields we also took 550-nm images. The mean exposure time at each wavelength was 3,400 s. In each field, we used standard software²⁴ to identify object positions and fluxes on the 650-nm image (which is the deepest image in each field), yielding roughly 150,000 objects per field. We have confirmed the robustness of the weighted intensity moments in our detected object catalogues by using different (FOCAS²⁵ with adaptive circular kernel²⁶) detection and evaluation software. At the position of each object, we evaluated the weighted moments at each wavelength, retaining only the measurements which the iterative weighted-moment algorithm did not flag as suspect. Measurements with small sizes ($I_{xx} < 1.0$ or $I_{yy} < 1.0$) were also excluded as suspect. The result is a list of multiple independent ellipticity measurements (and corresponding estimated measurement errors σ_i) for each object.

We then computed the best estimate of each galaxy's ellipticity by averaging the remaining measurements, weighted inversely by their estimated errors. If either of the ellipticity components at any wavelength deviated from this mean by more than $3\sigma_i$, that wavelength was eliminated and the process repeated. This step thus eliminates individual galaxy ellipticity measurements at wavelengths at which objects were noisy or blended, and it also reduces the systematic errors because the images at different wavelengths do not share the same residual point-spread function anisotropy. Finally we rejected objects with $\epsilon > 0.6$ as likely to be blends of more than one object, and applied flux density criteria ($1.8 \mu\text{Jy} > F_r > 0.11 \mu\text{Jy}$ through the 650-nm filter, 23–26 R-band magnitude) to select objects likely to be distant galaxies. We use these same selection criteria in calibrating the typical redshift of the background galaxies (below). The final catalogues contain about 45,000 galaxies in each field. A visual inspection of the final catalogues indicates that they are free of spurious objects such as bits of scattered light around bright stars.

These observed ellipticities must be corrected for the overall broadening effect of the point-spread function, which makes galaxies with elliptical profile appear more circular even if the point-spread function itself is perfectly isotropic. To calibrate this effect, we took a deep image with a very small point spread (the Hubble Deep Field South) and convolved it to the resolution of our final images, which is 1.07–1.25 arcsec as measured by the full-width at half-maximum at 650 nm. (The resolution on individual exposures, or ‘seeing’, was better, but the stellar size is larger in the final image with systematic shape errors removed from the point-spread function.) While some isolated galaxies became broader and less elliptical as predicted, most merged with their neighbours,

producing many more elliptical objects than predicted and preventing the construction of a clear relationship between observed and true ellipticity for individual galaxies.

Instead of constructing such a relationship we calibrated the fraction of cosmic shear recovered, as a function of resolution, from the ensemble of galaxies matching our selection criteria. We induced a known shear into the Hubble Deep Field South, convolved to the desired resolution, applied the same galaxy measurement and selection routines (at 650 nm only), and measured the mean ellipticity of the resulting sample. We averaged over repeated shears in several different directions to assess the measurement errors. The ratio of induced to recovered ellipticity was 4.5 ± 0.5 , with no clear trend as a function of resolution. The lack of such a trend would be quite surprising for isolated galaxies, but the coalescence of galaxy images appears to be the dominant effect. In the fairly small range of 1.07–1.25-arcsec resolution, this effect does not change the recovery factor by more than the measurement error of 0.5, so we adopt 4.5 as an overall ellipticity recovery factor.

Ellipticity correlations of distant galaxies

Two physically revealing ellipticity correlation functions have been defined¹⁰. In this approach, the ellipticity components of a galaxy i are calculated not with respect to the arbitrary x and y axes of the image, but with respect to the line joining it to another galaxy j (Fig. 3). Averaging over all galaxies i and j separated by angle θ on the sky, the correlations $\xi_1(\theta) \equiv \langle e_{1i}e_{1j} \rangle$ and $\xi_2(\theta) \equiv \langle e_{2i}e_{2j} \rangle$ have a unique signature in the presence of gravitational lensing, explained in detail in Fig. 3 legend. We recently reported the detection of a cosmic shear signal in the quadrature sum of these correlation functions²⁷.

Figure 4 shows the ellipticity correlations for each of the three fields in the angular separation range 2–36 arcmin (top panels). The

plotted errors indicate 68% confidence intervals determined from 200 realizations (bootstrap-resampled) of the final galaxy catalogue in each field. Note that the measurements in different angular separation bins are not statistically independent, but ξ_1 and ξ_2 are independent from each other, as are the three fields. At $\theta = 6.1$ arcmin, the confidence that $\xi_1 > 0$ is 97%, 99.5% and 99.5% for the three fields in the order shown in Fig. 4. Similarly, the confidence that $\xi_2 > 0$ at the same angular scale is 87.5%, >99.5% and 97% respectively. Some cosmic variance, that is, real systematic differences among fields of this size is expected²⁸, but the statistical errors are too large to examine this effect. We plot the average over the three fields in the lower panels of Fig. 4, with 1σ errors in the mean derived from the variance among the fields (black points and errors). The signature of gravitational lensing by large-scale structure is evident: ξ_1 declines as the angular scale increases, but is positive at all scales, while ξ_2 matches ξ_1 at small scales but drops below zero at large scales. This result is robust. Similar, but lower signal-to-noise, profiles are obtained if we use unweighted moments or moments from the 650-nm images only.

We performed several tests for systematic errors. The effects of residual point-spread function anisotropy are demonstrated by plotting the correlation functions of the stars (blue in Fig. 4). These are far closer to zero than are the galaxy correlations. Only ξ_1 in the innermost bin has an apparently significant stellar correlation. To test the effect that this might have on the galaxy correlations, we computed the star–galaxy correlations $\langle e_{1,\text{star}}e_{1,\text{gal}} \rangle$ and $\langle e_{2,\text{star}}e_{2,\text{gal}} \rangle$ (green in Fig. 4). The star–galaxy correlations are extremely close to zero in this bin. There are also tests involving the galaxy sample alone. The cross-correlation $\xi_3 \equiv \frac{1}{2}(\langle e_{1i}e_{2j} \rangle + \langle e_{2i}e_{1j} \rangle)$ should vanish in the absence of systematic errors (red in Fig. 4). The result is reassuringly close to zero. The plotted errors for ξ_3 can also

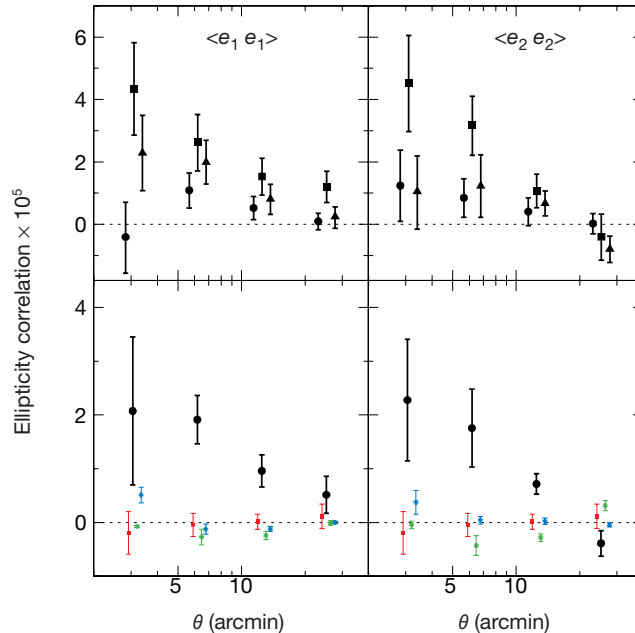


Figure 4 Detection of ellipticity correlations. The upper panels show the measured ellipticity correlations as a function of angle for three independent fields covering a total of 1.5 square degrees (ξ_1 at left and ξ_2 at right). Markers have been slightly offset horizontally for clarity. From left to right in each bin are fields at 11 h 38 min, $-12^\circ 33'$, 23 h 48 min, $+00^\circ 57'$, and 04 h 29', $-36^\circ 18'$ (J2000). In each field, roughly 45,000 faint galaxies passed all the filters and significance tests, from an initial catalogue of about 150,000 objects. Errors shown are 68% confidence intervals determined from 200 bootstrap-resamplings of the galaxy catalogues. The lower panels show the mean of the ellipticity correlation functions over the three fields (black), with 1σ errors derived from the variance between fields. The behaviour as a function of angle matches that expected from weak gravitational lensing by large-scale structure. The lower panels also contain several

null tests of systematic error. The cross-correlation of the galaxies ξ_3 should vanish in the absence of systematic error, and in fact is everywhere consistent with zero (red). The ellipticity correlations of stars (blue) are everywhere consistent with zero except in the innermost bin of ξ_1 . The effect of non-zero stellar correlations on the galaxy correlations is illustrated by the star–galaxy correlation (green), which is very close to zero in this bin. An additional test of systematics, a search for preferential alignment of galaxies with the CCD axes, is also null. Though galaxy ellipticity correlations continue to rise at smaller angles, the smaller number of galaxy pairs makes the measurement noisier, there are few closely-spaced stars to assess systematic error, and the theoretical interpretation on small scales is difficult.

be taken as an indicator of the statistical error associated with the number and distribution of galaxies included in the catalogues (but reduced due to the averaging of two functions in ξ_3). This estimate of statistical error agrees roughly with that shown for ξ_1 and ξ_2 . Finally, the weak lensing signature disappears if we randomize the galaxy positions.

Apart from these null tests, there are also affirmative tests. One test is to take similar data centred on a cluster of galaxies of known mass. A 650-nm image of massive cluster at redshift 0.45, taken with the same camera and processed in the same way, exhibits correlation functions (ξ_1 and ξ_2) of the expected angular dependence and of larger amplitude than in any of the three blank fields, despite likely contamination of the galaxy sample by cluster members. ξ_3 also vanishes in this field. Another test involves inverting the back-

ground galaxy ellipticity distribution to yield a map of projected mass in each blank field. We find occasional mass concentrations which can often be identified with likely foreground clusters, but no linear features or pile-ups at the edges of the image which might indicate problems in the background galaxy catalogues. Furthermore, when mass is mapped using only those galaxies likely to be behind a serendipitous cluster (on the basis of colour information), the lensing signal from that cluster increases markedly. This corroborates the idea that the correlation functions are accumulating over many sources and many overdensities spread throughout the line of sight. All these tests indicate that we have indeed measured cosmic shear in our blank fields and that contamination from surviving systematic error is low. We now turn to comparisons with theoretical predictions of this effect.

Comparison with theoretical predictions

Correlations between the ellipticity of galaxies increase strongly with background galaxy redshift, so we must first constrain the source redshift distribution, $N(z)$. Very little is known about the redshift distribution of galaxies as faint as those used here, so we assume a simple model where $N(z) \propto z^2 \exp(-z/z_0)$, and adjust z_0 to match weak gravitational lensing observations of a high-redshift galaxy cluster of known velocity dispersion (MS1054 at $z = 0.83$)²⁹. We observed this cluster with the same camera and telescope and reduced the data in the same way as for the blank fields, and compared the faint galaxy ellipticities (tangential to the cluster centre) to that expected for a range of z_0 . We found that $z_0 = 0.5$ was the best match.

This model of $N(z)$ was used as input to a cold-dark-matter simulation code developed by Hu and Miralda-Escudé. This code³⁰ computes the shear power spectrum and correlation function for any given cosmology, using the prescription of refs 31 and 32. From this, the mass power spectrum is calculated in the nonlinear regime when the growth of dark-matter structures which have gravitationally collapsed, has modified the mass spectrum. Results were obtained for three cosmological models, and are plotted along with our seeing-corrected measurements on a logarithmic scale in Fig. 5. Two current models were normalized to the microwave background fluctuations (COBE) at large angle and to local galaxy cluster abundance (assuming that mass traces light) at small angle: an open universe with $\Omega_{\text{matter}} = 0.45$ (orange in Fig. 5), and a flat universe dominated by a cosmological constant $\Lambda = 0.67$ (green, solid line). The agreement between the data and the two viable cosmological models is impressive for a first measurement. For comparison purposes we also show the old standard cold dark matter flat cosmology (blue), which is only normalized to COBE. (A full listing of the parameters used in these models is shown in Table 1.) To illustrate the effect of varying $N(z)$, we also plot the Λ -dominated cosmology with $z_0 = 0.3$ (green, dotted line). Since our modelled $N(z)$ peaks at $z = 2z_0$, this lowers the typical redshift from 1.0 to 0.6. Decreasing z_0 decreases the amplitude of the correlations, but has little effect on their shapes. The uncertainty in $N(z)$ implies a factor of several uncertainty in the amplitude of the correlation, and is by far the dominant calibration error.

Despite this uncertainty, standard cold dark matter normalized to COBE is ruled out by the measured values of ξ_1 . Although this is not surprising, it is the first (to our knowledge) cosmological constraint from wide-field weak lensing and it agrees with several other methods which disfavour this model^{33,34}. The other two models are consistent with the data at the 3σ level. This indication of a low- Ω_{matter} universe here is in agreement with a remarkable array of independent methods, including type Ia supernovae, cosmic microwave background anisotropies, cluster baryon fraction together with cluster mass (lensing) and primaeval deuterium, and the age of the oldest stars coupled with the Hubble constant³⁵. However, the shape of ξ_2 is not a good fit to either of these two model cosmologies, which are based on a single power-law mass spectrum.

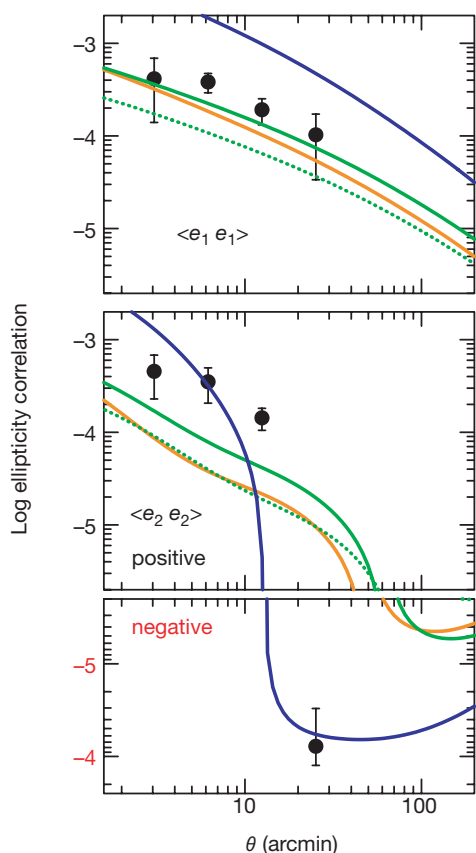


Figure 5 Comparison of ellipticity correlations with predictions. We plot our measurements of ellipticity correlations with 1σ errors on a logarithmic scale along with theoretical predictions based on various models for a cold-dark-matter universe. The top panel shows ξ_1 , the middle panel shows $\xi_2 > 0$, and the lower panel shows $\xi_2 < 0$. In each panel, the blue theoretical curve is for the standard cold dark matter model, the solid green curve is for a universe with a cosmological constant (Λ CDM), and the orange curve is for an open universe; detailed parameters are given in Table 1. The dotted green curve shows the effect of decreasing the mode of the background galaxy redshift distribution, $2z_0$, from 1.0 to 0.6 for one model (Λ CDM). The errors shown are derived from the variance among three fields. The data from Fig. 4 have been multiplied by a correction factor of 20 here, to compensate for the ellipticity dilution factor of 4.5 described in the text, which is squared in the correlation functions. The measurements are consistent with Λ CDM and an open universe at the 3σ level despite the visual impression given by ξ_2 , which is due to the logarithmic axes. Standard cold dark matter is inconsistent with ξ_1 at many sigma. This measurement of ellipticity correlations due to cosmic shear over half-degree angular scales is in agreement with a variety of other evidence in ruling out standard cold dark matter. Weak lens observations of larger fields and more distant galaxies will be able to clearly distinguish between the remaining models, or suggest the need for a new model.

If confirmed by further data, this would suggest the need for a more complicated mass spectrum.

This technique can further distinguish between open and Λ -dominated universes if extended to the somewhat larger angular scales where those cosmologies predict ξ_2 will cross zero as shown in Fig. 5. A survey of many $2^\circ \times 2^\circ$ fields now underway will rule out one or more of these cosmologies at the $\sim 8\sigma$ level at 10-arcmin angles (3σ level for a differential measure of the slope of the power spectrum). Separating the background galaxies into discrete redshift bins based on multi-colour photometry will enable measurement of the ellipticity correlation (or equivalently the dark-matter power spectrum) as a function of cosmic time; wide-field weak lensing surveys deep enough to identify galaxies at $z \sim 2$ and measure their shapes will constrain several cosmological parameters³⁰. Ultimately, the combination of all the power-spectrum probes (lensing, cosmic microwave background, galaxy distributions, and peculiar velocities) will tightly constrain theories of the origins of fluctuations in the early Universe and their growth into galaxies and large-scale structure. $\square$

Received 13 December 1999; accepted 21 March 2000

1. Page, L. & Wilkinson, D. T. The cosmic microwave background. *Rev. Mod. Phys.* **71**, 173–179 (1999).
2. Tyson, J. A., Valdes, F. & Wenk, R. Detection of systematic gravitational lens galaxy image alignments: mapping dark matter in galaxy clusters. *Astrophys. J.* **349**, L1–L4 (1990).
3. Fahlman, G., Kaiser, N., Squires, G. & Woods, D. Dark matter in MS1224 from distortion of background galaxies. *Astrophys. J.* **437**, 56–62 (1994).
4. Squires, G., Kaiser, N., Fahlman, G., Babul, A. & Woods, D. A weak gravitational lensing analysis of Abel 2390. *Astrophys. J.* **469**, 73–77 (1996).
5. Clowe, D., Luppino, G. A., Kaiser, N., Henry, J. P. & Gioia, I. M. Weak lensing by two $z \sim 0.8$ clusters of galaxies. *Astrophys. J.* **497**, 61–64 (1998).
6. Hoekstra, H., Franx, M., Kuijken, K. & Squires, G. Weak lensing analysis of CL 1358+62 using hubble space telescope observations. *Astrophys. J.* **504**, 636–660 (1998).
7. Mellier, Y. Probing the universe with weak lensing. *Annu. Rev. Astron. Astrophys.* **37**, 127–189 (1999).
8. Gunn, J. E. A fundamental limitation on the accuracy of angular measurement in observational cosmology. *Astrophys. J.* **147**, 61–72 (1967).
9. Dyer, C. & Roeder, R. Observations in locally inhomogeneous cosmological models. *Astrophys. J.* **189**, 167–175 (1974).
10. Miralda-Escudé, J. The correlation function of galaxy ellipticities produced by gravitational lensing. *Astrophys. J.* **380**, 1–8 (1991).
11. Blandford, R., Saust, A., Brainerd, T. & Villumsen, J. The distortion of distant galaxy images by large scale structure. *Mon. Not. R. Astron. Soc.* **251**, 600–627 (1991).
12. Kaiser, N. Weak gravitational lensing of distant galaxies. *Astrophys. J.* **388**, 272–286 (1992).
13. Villumsen, J. Weak lensing by large-scale structure in open, flat and closed universes. *Mon. Not. R. Astron. Soc.* **281**, 369–383 (1996).
14. Jain, B. & Seljak, U. Cosmological model predictions for weak lensing. *Astrophys. J.* **484**, 560–573 (1997).

15. Kaiser, N. Weak lensing and cosmology. *Astrophys. J.* **498**, 26–42 (1998).
16. Kristian, J. On the cosmological distortion effect. *Astrophys. J.* **147**, 864–867 (1967).
17. Valdes, F., Tyson, J. A. & Jarvis, J. F. Alignment of faint galaxy images: cosmological distortion and rotation. *Astrophys. J.* **271**, 431–441 (1983).
18. Mould, J. et al. A search for weak distortions of distant galaxy images by large-scale structure. *Mon. Not. R. Astron. Soc.* **271**, 31–38 (1994).
19. Schneider, P. et al. Detection of shear due to weak lensing by large-scale structure. *Astron. Astrophys.* **333**, 767–778 (1998).
20. Wittman, D. et al. Big SPIE throughput camera: the first year. *Proc. SPIE* **3355**, 626–634 (1998).
21. Castleman, K. R. *Digital Image Processing* 214 (Prentice Hall, Englewood Cliffs, New Jersey, 1979).
22. Fischer, P. & Tyson, J. A. The mass distribution of the most luminous x-ray cluster RXJ1347.5-1145 from gravitational lensing. *Astron. J.* **114**, 14–24 (1997).
23. Gullixson, C. A., Boeshaar, P. C., Tyson, J. A. & Seitzer, P. The *B/RI* photometric system. *Astrophys. J. Suppl. Ser.* **99**, 281–293 (1995).
24. Bertin, E. & Arnouts, S. SExtractor: software for source extraction. *Astron. Astrophys. Suppl.* **117**, 393–404 (1996).
25. Valdes, F. Resolution classifier. *Proc. SPIE* **331**, 465–472 (1982).
26. Tyson, J. A. in *AIP Conf. Proc. Dark Matter* (eds Holt, S. & Bennett, C.) 287–296 (AIP Press, 1995).
27. Wittman, D. & Tyson, J. A. The shear correlation function out to 20 arcminutes. In *Gravitational Lensing: Recent Progress and Future Goals* (eds Brainerd, T. G. & Kochanek, C. S.) (ASP Conference Series, Astronomical Society of the Pacific, San Francisco, in the press).
28. Kruse, G. & Schneider, P. The non-gaussian tail of cosmic shear statistics. Preprint astro-ph/9904192 at (<http://xxx.lanl.gov>) (1999).
29. Tran, K. H. et al. The velocity dispersion of MS1054-03: a massive galaxy cluster at high redshift. *Astrophys. J.* **522**, 39–45 (1999).
30. Hu, W. Power spectrum tomography with weak lensing. *Astrophys. J.* **522**, L21–L24 (1999).
31. Hamilton, A. J. S., Matthews, A., Kumar, P. & Lu, E. Reconstructing the primordial spectrum of fluctuations of the universe from the observed nonlinear clustering of galaxies. *Astrophys. J.* **374**, L1–L4 (1991).
32. Peacock, J. A. & Dodds, S. J. Non-linear evolution of cosmological power spectra. *Mon. Not. R. Astron. Soc.* **280**, L19–L26 (1996).
33. Ostriker, J. P. & Steinhardt, P. J. The observational case for a low density universe with a cosmological constant. *Nature* **377**, 600–602 (1995).
34. Bahcall, N. A., Ostriker, J. P., Perlmutter, S. & Steinhardt, P. J. The cosmic triangle: revealing the state of the universe. *Science* **284**, 1481–1488 (1999).
35. Turner, M. S. & Tyson, J. A. Cosmology at the millennium. *Rev. Mod. Phys.* **71**, 145–164 (1999).
36. Frenk, C., White, S. D. M., Efstathiou, G. & Davis, M. Galaxy clusters and the amplitude of primordial fluctuations. *Astrophys. J.* **351**, 10–21 (1990).

Acknowledgements

We thank W. Hu and J. Miralda-Escudé for help with theoretical predictions of several cosmological models. We also thank S. Gentile for artwork, and the staff of CTIO for their help with the BTC project and for their upgrading and maintenance of the delivered image quality of the Blanco telescope. Cerro Tololo Inter-American Observatory is a division of National Optical Astronomy Observatory (NOAO), which is operated by the Association of Universities for Research in Astronomy, Inc., under cooperative agreement with the NSF. Big Throughput Camera construction was partially supported by the US National Science Foundation.

Correspondence and requests for materials should be addressed to D.M.W. (e-mail: wittman@physics.bell-labs.com).

Prestin is the motor protein of cochlear outer hair cells

Jing Zheng*, Weixing Shen*, David Z. Z. He*, Kevin B. Long†, Laird D. Madison† & Peter Dallos*

* Auditory Physiology Laboratory (The Hugh Knowles Center), Departments of Neurobiology and Physiology and Communication Sciences and Disorders, Northwestern University, Evanston, Illinois 60208, USA

† Center for Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Northwestern University Medical School, Chicago, Illinois 60611, USA

The outer and inner hair cells of the mammalian cochlea perform different functions. In response to changes in membrane potential, the cylindrical outer hair cell rapidly alters its length and stiffness. These mechanical changes, driven by putative molecular motors, are assumed to produce amplification of vibrations in the cochlea that are transduced by inner hair cells. Here we have identified an abundant complementary DNA from a gene, designated *Prestin*, which is specifically expressed in outer hair cells. Regions of the encoded protein show moderate sequence similarity to pendrin and related sulphate/anion transport proteins. Voltage-induced shape changes can be elicited in cultured human kidney cells that express prestin. The mechanical response of outer hair cells to voltage change is accompanied by a 'gating current', which is manifested as nonlinear capacitance. We also demonstrate this nonlinear capacitance in transfected kidney cells. We conclude that prestin is the motor protein of the cochlear outer hair cell.

Cochlear hair cells are non-neuronal epithelial cells that transduce acoustic signals. Outer hair cells (OHCs) are responsible for the exquisite sensitivity and frequency-resolving capacity of the normal mammalian hearing organ¹; they provide local mechanical amplification (the 'cochlear amplifier') in the form of feedback. In contrast, inner hair cells convey auditory information to the brain. Outer hair cells have cylindrical somata of constant diameter and variable length. When their membrane potential is altered^{2–5}, somatic shape changes of up to 5% occur; the cell shortens when depolarized and lengthens when hyperpolarized^{13,5}. Length changes

do not depend on either ATP or Ca²⁺ (ref. 6), and they can be elicited with unchanging amplitude at microsecond rates up to high audio frequencies^{7,8}. Motile responses are accompanied by charge movement, which is reflected in nonlinear capacitance^{9,10}, akin to the translocation of gating charges of voltage-gated ion channels¹¹. This nonlinear capacitance is widely used as a 'signature' of the electro-motile process^{10,12}. Motility is also accompanied by a change in the axial stiffness of the cell¹³. By virtually any test, electromotility and electrically-induced stiffness changes are correlated (D.Z.Z. and P.D., unpublished data) and we collectively describe them as

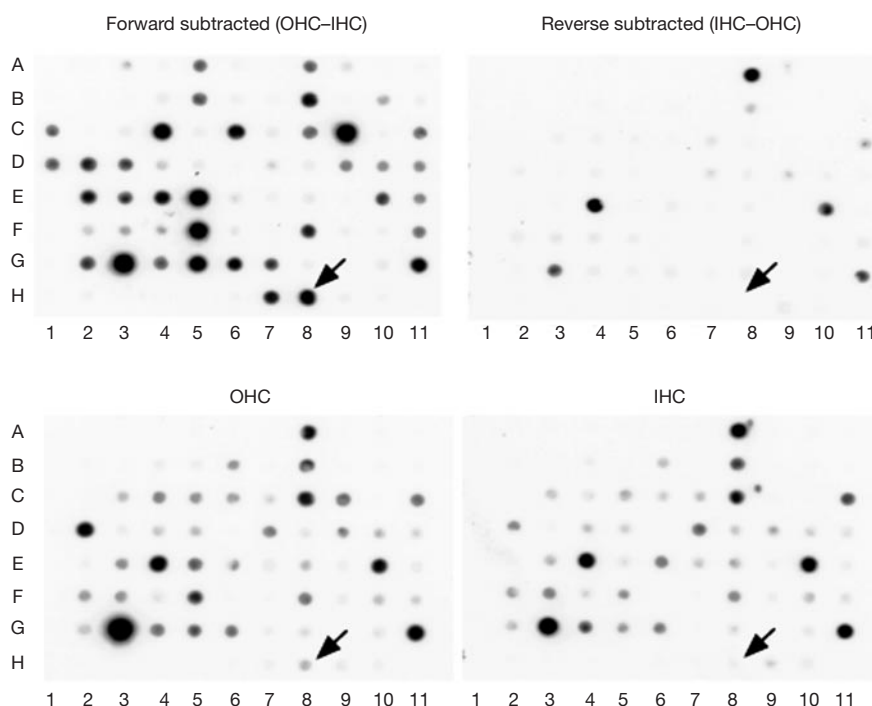


Figure 1 *Pres* fragment identification. Four identical cDNA blots of PCR-amplified inserts derived from the subtracted cDNA library were separately hybridized with radioactive forward-subtracted probe (OHC–IHC), reverse-subtracted probe (IHC–OHC), and OHC

and IHC cDNA probes. Note, for example, that the samples at the H8 locations (arrows) show strong signals in OHC and OHC–IHC hybridization blots and little or no signal in the IHC and IHC–OHC blots. Samples such as these were selected for sequencing.

voltage-dependent mechanical changes of the OHC, hereafter called 'electromechanics'. These observations make it apparent that the fast mechanical changes in OHCs are powered by a molecular motor that is fundamentally different from other biological force generators, such as the myosin, kinesin or dynein families. The OHC molecular motor performs direct, rapid, reversible electro-mechanical conversion.

Outer hair cell electromotility is the probable result of the concerted action of a large number of independent molecular motors that are closely associated with the cell's basolateral membrane^{6,14,15}, possibly by the densely packed 10-nm particles¹⁶ seen therein. A general hypothesis is that the particles are the molecular sensor motors and here we attempt to identify them. There have been some suggestions as to the identity of these motor molecules. On the basis of similarities between the cortical structures of erythrocytes and OHCs, the motor molecule has been proposed to be a modified anion exchanger^{17,18}. Because their shallow voltage dependence matches that of charge movement in OHCs, transporters have been favoured, as opposed to modified voltage-dependent channels, as likely candidates¹². One suggestion has been that the motor is related to a fructose transporter, GLUT5 (ref. 19). Until now, molecular identification of the motor protein has not been achieved. In as much as the most distinguishing feature of this molecular motor is its speed, we designate it as prestin, from the musical notation *presto*. The gene (*Prestin*) coding for this protein is abbreviated *Pres*.

Cloning of *Pres* and the properties of prestin

Our molecular approach to identify the OHC sensor-motor protein faced two initial challenges: first, isolation of complementary DNA from limited amounts of biological material; and second, recognition of the candidate motor protein cDNA among the isolated unknown genes. Isolation of a sufficient number of OHCs is a fairly routine task. Obtaining comparable numbers of inner hair cells (IHCs), however, required some modification of techniques (D.Z.Z.H. *et al.*, unpublished data). We were able to isolate messenger RNA from about 1,000 OHCs and 1,000 IHCs and used mRNA reverse transcription followed by 5'-mRNA cap and oligo dT-dependent PCR amplification to obtain OHC and IHC cDNA pools.

Our strategy for recognizing the cDNAs that might encode the candidate motor protein was based on the following assumptions:

the protein is expressed in OHCs and not in the non-motile IHC; ontogenic expression should relate to the known development of electromechanical responses²⁰; the protein should be relatively abundant in OHCs; the protein should be a transmembrane protein; and electromechanical responses should be demonstrable when the protein is expressed in a heterologous system. We expected that gene expression related to mechano-electric transduction and genomic, metabolic and structural functions would be shared by these two types of hair cell; therefore, we applied a suppression subtractive hybridization polymerase chain reaction (PCR) procedure to amplify and enrich the OHC cDNA pool for uniquely expressed gene products²¹. This procedure uses two rounds of cDNA hybridization to subtract out common OHC cDNAs using a vast excess of IHC cDNA. The unique fragments are amplified by PCR, digested with *RsaI* and cloned to create a library of candidate clones. This technique is useful because it simultaneously enriches differentially expressed cDNA fragments and suppresses non-target DNA amplification.

About 1,300 fragment clones were screened with four cDNA pools (OHC-IHC, IHC-OHC, OHC unsubtracted and IHC unsubtracted) to identify differentially expressing clones (Fig. 1). Once false positives were excluded, 487 clones were identified as differentially expressing, and of these 108 were sequenced. Because of duplicates, these 108 sequences identified 50 unique cDNA fragments. Eighteen of the clones had sequences highly similar to known proteins, whereas 32 were thought to be unique in that there were no obvious homologous sequences in the genetic database. Eleven of the 32 clones had apparent open reading frames and were examined in greater detail by dot-blot hybridization with radioactively labelled cloned cDNA against immobilized OHC and IHC cDNA pools. *Pres* was one of the 11 candidate clones whose differential expression was most consistent with our assumptions about the motor protein. *Pres* fragments in the PCR subtracted library were abundant (>10% of 487 differentially expressed clones) and showed consistent differential hybridization with OHC- and IHC-derived probes (Fig. 1). Using one of these *Pres* cDNA *RsaI* fragments as a probe, we subsequently isolated three cDNA clones from a λ gt11 gerbil adult cochlea library (provided by B. Schulte). The largest of these clones was 4.1 kilobase (kb) with an open reading frame of 2,232 base pairs (bp), encoding a 744-amino-acid protein (Fig. 2). The sequence around the putative translation start site contains a consensus Kozak sequence²² and an in-frame

```

gPrestin 1 ...MDHAENEIEPVATQKVERPIFISHPVLQERLEVKDKVSESIGDKLQAFCTCTPKKIRNIYMFLLPITKWLPAKYKFKEYVLGDLVSGISTGVQLP
rPendrin 1 MAARRRSEPPQLAEYSCSYAVSRPVYSELAFQQQRERRLPERRTLRSLARSCSCSRKRAFALKALLPILDWLPKYRVKEWLLSDIISGVSTGLVGT

97  QGLAFAMLAAYEPVFLGYSSFFVIMYCFEGTSHHISIGPFAVISLMIGGVAVRVDD. IVIPGG...VNATN. GTEARDALRVKVMASVTLLSGIIQ
101 QGMAYALLAAVFPQYGLYSAFFPILTYFVFGTSHHISVGPFPVSLMVGSVLSMAPDDHFLVPSGNGSTLNTTLDTGTRDAARVLLASTLTLVGLIIQ

191 FCLGVCRFGFVAIYLTEPLVRGFTTAAAVHFTSMKYLFGVKTARYSGIFSVYXSTVAVLQNVKNLNVCSLGVGLMVFGLLLGGKEFNERFKEKLPAPI
201 LVFGGLQIGFIVRYLADPLVGGFTTAAAFQVLVSQKLIVLNVSTKNYNGVLSIIYTLIEIFQNIQIGDTNIADFIAGLLTIIVCMVAKELNDRFKHKIPVPI

291 PLEFFAVVMGTGISAGFNHESYSVDVVGTLPLGLLPPANPDTSFLHLYVVDIAIAIVGFSVTISMATLANKHGYQVDGNQELIALGICNSIGSLFQT
301 PIEVIVTIIATAISYGANLEANYNAGIVKSIPIPSGFLPPVLPVSVGLFSDMLAASFSAIVVAYAIASVGVKYATKHDYIIDGNQEFIAFGISNVFSGGFSC

391 FSISCSLSRSLVQEGTGKQTLAGCLASLMILLVILATGFLFESLPQAVLSAIVIVNLKGMFMQFSDLPFFWRTSKIELTIWLTTFVSSSLFLGLDYGLIT
401 FVATTALSRATVQESTGGKTQVAGLISAVIMVAIVALGKLEPLQKSVLAADVIANLKGMMQVCDVPLRWKQNKTDVAVIWFVFTCMSIILGLDLGLLA

491 AVIIALLTVIYRTQSPSYKVLGQLPDTDVYIDIDAYEEVKEIPGKIFQINAPIYVANS DLYSNALRKRTGVNPAALIMGARRKAMRYKAEVGNANIANA
501 GLLFGLLTVVLRVQFPWNGLSVPSDTIYKSI THYKNLEEPEGVKILRFSSPIFYGNVDGFKCKVSTVGFDAIRVYNKRLKALRRRIQKLIKKGQLRAT

591 AVVKV DGEVDGENATKPE...EEDDEVKYP...PIVIKTTFPEEL...QRMPQNTENVHTIILDFTQVNFIDSVGVKTLAVMVKYGDVGIYVYLAGCSPQV
601 KNGIISDVGSSNNAFEPDEDEVPEELDIPTKEIEIQVDWNSLPVKVNVKVP. IHSVLVDCGAVSFLDVGVRSRLRMIVKEFQRIDVNVYFALLQDDV

684 VNDLTRNRFFENPALKELLFHSIHDAVLGSHVREAMEAQEASAPPPQDDMEPNATPTTPEA 744
700 LEKMEQCQGFDDNIRKDRFELTVHDAIL.....YLQNAKSREGQDSLLETITLIQDCKDPELMEABINEEELDVQDEAMRRLAS 780

```

Figure 2 Amino-acid sequence of gerbil prestin and its homology to rat pendrin. Identical prestin/pendrin residues are shown in red, and highly similar amino acids are in green. Prestin residues, which are in bold and underlined, indicate the 12 differences for the deduced human sequence derived from exons 1–6 of the known *PRES* genomic

sequence (first 295 residues). The position of the signature sulphate transport motif is boxed (black). A blue box indicates the positive charge cluster and a yellow box denotes the negative-charge cluster. GenBank accession numbers: gerbil prestin, AF230376; rat pendrin, AF167412.

stop codon is present 12 bases upstream of this predicted start site. There is a short 223-bp untranslated 5' end and a large 1,654-bp, 3' untranslated region. A computer search with the *Pres* sequence revealed that about one-third of the human *PRES* gene had been sequenced as part of the human chromosome 7 effort (BAC clone RG107G13, bases 99,895 to >113,558). The amino-acid homology between human and gerbil prestin, deduced from the genomic sequence of the first 6 exons, is 98% (Fig. 2).

Analysis of the prestin sequence revealed that its highest homology was to members of a family of sulphate/anion transport proteins including pendrin and DRA (down-regulated in adenoma)^{23,24}. The pendrin homology (40%) was of particular interest because this chloride–iodide transport protein is a cause of the genetically inherited deafness observed in Pendred's syndrome²⁵. This homology, although modest overall, is similar to the homology among other members of the sulphate/anion transport family. Because we were using gerbil cDNA, and only rat, mouse and human pendrin sequences were known at the time, it was necessary to prove that prestin was not the gerbil homologue of pendrin; first it should be noted that mouse pendrin is expressed in the endolymphatic duct and sac, the utricle and saccule and in the external sulcus but not in the organ of Corti²⁶ in the cochlea; second, we cloned an 824-bp fragment of gerbil pendrin cDNA, the sequence of which was distinct from that of prestin (data not shown). The gerbil pendrin amino-acid sequence derived from the above cDNA fragment was 88% homologous to that of human pendrin, but only 38% homologous to that of gerbil prestin.

The protein prestin (744 amino acids; relative molecular mass (M_r) 81.4 K) is hydrophobic, with ~50% non-polar residues; 27.8% of the protein is composed of the amino acids valine, leucine and isoleucine. Computer modelling of the amino-acid sequence produces ambiguous results using multiple modelling programs. Specifically, the models TMHMM²⁷ and TMPred²⁸ report ambiguous results as to the number and location of transmembrane regions and are unable to predict the topology of the amino and carboxy termini in relation to the membrane. In comparison, when pendrin is subjected to the same modelling programs, 11 transmembrane regions and an intracellular N terminus are unambiguously predicted. A complete topology analysis using antibodies directed against prestin epitopes, affinity labelling and mutagenesis will probably be required to clarify the nature of the protein/membrane interface. Two distinctive charged regions are located in the C-

terminal region. A positive-charge cluster is located at residues 557–580; adjacent to this is a negative-charge cluster at residues 596–613 (Fig. 2). Prestin's overall predicted hydropathy profile is similar to those of pendrin, DRA and other members of the sulphate/anion transport family. The homology to pendrin is highest in the 50-amino-acid region (97–146) that encompasses the sulphate transport motif (109–130), but does not have long regions of identity elsewhere. Instead, there is a pattern of moderate homology by both amino-acid identity and similarity throughout the protein with multiple, discrete, 8–12-residue segments of identity. The charged cluster regions of prestin and pendrin are conserved in their net charges, but are otherwise not remarkably homologous. The proteins differ most in the region between the two charge clusters and at the termini. Prestin shows high homology in the sulphate transport region, a highly conserved domain found in homologues in organisms as distant as yeast, *Caenorhabditis elegans* and plants. However, both human and gerbil prestin differ from the consensus sequence at three positions (Fig. 3)²⁹. Consequently, although prestin appears to be related to sulphate transporters, these differences in the conserved region suggest that the protein may have distinct properties.

We then determined the tissue-specific and developmental expression pattern of *Pres* in the gerbil. Northern analysis with a *Pres* probe did not detect expression in liver, lung, brain, spleen, ovary, kidney, muscle or heart (total RNA 20 µg; data not shown). Virtual northern dot-blot analysis using cDNA prepared from IHC, OHC, mature and newborn organs of Corti, and thyroid revealed expression only in mature OHCs and 20-day-old organs of Corti (Fig. 4a). Similar results were obtained using *Pres*-specific primers and a qualitative PCR assay (Fig. 4b). Both virtual northern and PCR showed progressively increasing *Pres* expression in isolated organ of Corti from birth up to 20 days after birth (Fig. 4c, d), by which time electromotility is fully developed in the altricial gerbil²⁰. In OHCs, the motors are generally thought to be congruent with abundant ~10-nm membrane particles^{16,30}. The density of these particles in gerbil OHCs shows a developmental increase, decelerating towards adult values 16–18 days after birth³¹. Thus, the ontogenic expression of *Pres* is similar to

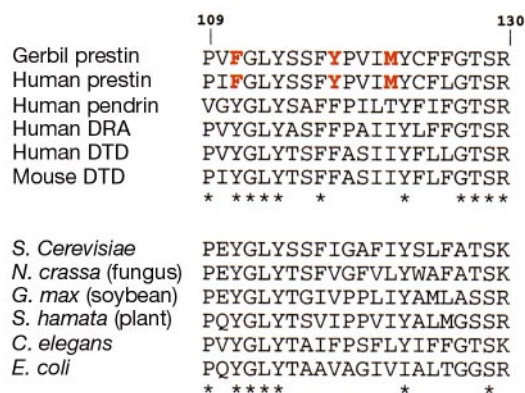


Figure 3 Comparison of sulphate transport motifs. The amino-acid sequence of gerbil and human prestin is compared with the same domain in other members of the sulphate/anion transport family. Top panel, alignment of prestin, pendrin, DRA and DTD (diastrophic dysplasia) from human and rodent species. Bottom panel, sulphate transport motifs from a variety of putative transporters in lower organisms. Amino-acid residue numbers refer to the sequence of gerbil prestin. Invariant residues are indicated by the asterisks below the sequence panels. Similarly divergent residues in the two prestin sequences are coloured red. The sulphate transport motif is defined by PROSITE PS01130.

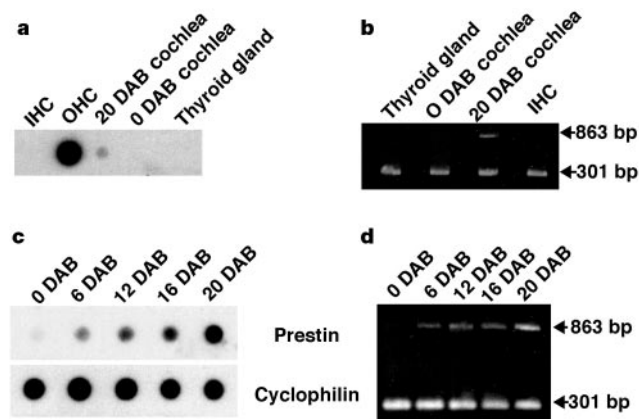


Figure 4 Analysis of *Pres* expression in SMART cDNA from different tissues and at different developmental ages. **a**, Virtual northern dot-blot analysis. cDNA pools derived from mRNA of gerbil tissues were hybridized with a ³²P-labelled *Pres* probe. **b**, PCR analysis. cDNA pools synthesized from different tissues were amplified with *Pres* and cyclophilin-specific primers. In **a** and **b**, 'cochlea' refers to organ of Corti-derived tissue. **c**, Virtual northern dot-blot analysis. cDNA pools derived from organ of Corti mRNA at 0, 6, 12, 16 and 20 DAB were hybridized with a ³²P-labelled *Pres*-specific probe. The blot was then stripped and rehybridized with a ³²P-labelled cyclophilin probe. **d**, PCR analysis. cDNA pools derived from organ of Corti mRNA at 0, 6, 12, 16 and 20 DAB were amplified with *Pres* and cyclophilin-specific primers. DAB, days after birth. The exposure time was shorter in **a** than in **c** because of the strong signal from OHC cDNA.

that of membrane particles, electromotility and the onset of high sensitivity hearing.

Functional tests of prestin

To investigate the function of the prestin cDNA, we sought to examine the functional properties of prestin in eukaryotic cells. To accomplish this, we subcloned *Pres* into the eukaryotic expression vector pcDNA3.1. We also created a C-terminal epitope-tagged (V5) version (pcDNA6/V5-HisA vector) to determine expression of the full-length protein in transfected cells. The expression of the V5 version was examined by indirect immunofluorescence to assure its presence in transfected cells, as we did not have an antibody directed against prestin. The epitope-modified prestin was apparently located in the cell membrane, with a punctate distribution in permeabilized cells (data not shown). In addition, western blot analysis using an anti-V5 antibody against transfected cell extracts showed production of the appropriate sized protein (data not shown).

To test the sensor–motor function of prestin, voltage-dependent charge movement was measured from TSA201 cells after transient transfection of unmodified, native *Pres* cDNA. A second plasmid (pEGFP-N2) containing the green fluorescent protein (GFP) cDNA was used as an independent marker for successful transfection of the cells. Transfection efficiency was 30–40% as judged from the fluorescence of GFP in transfected cells. Consistent with this, under voltage clamp, 36.8% of cells transfected with *Pres* alone displayed transient capacitive currents, using a standard subtraction protocol after blocking ionic currents (Fig. 5a). In experiments conducted on cells cotransfected with prestin and GFP cDNA, transient capacitive currents were obtained from 91.8% of the GFP-positive cells (45 out of 49). In contrast, untransfected cells ($n = 20$) or cells transfected with only the control plasmid ($n = 28$) had no measurable nonlinear capacitance. The capacitive currents were directed towards the outside as the membrane potential was depolarized and towards the inside as it was hyperpolarized. Charges transferred at command onset and offset were roughly equal. The decay of transient currents was rapid (100–200 μ s) and could be well fitted to a single exponential. The voltage protocol used to estimate membrane capacitance and charge transfer across the membrane was an ascending stair-step voltage waveform¹⁵ (Fig. 5b). The membrane capacitance was fitted to the derivative of a first-order Boltzmann equation (ref. 15; see Methods and Fig. 5c). In a group of seven cells, these curve fits yielded values of $Q_{\max} = 1.76 \pm 0.20$ pC, $V_{1/2} = 57.30 \pm 2.98$ mV, $1/\alpha = kT/ze = 28.00 \pm 2.13$ mV and $C_{\text{lin}} = 20.50 \pm 2.65$ pF, respectively (mean $\pm$ s.e.m.). The valence can be calculated from α , $z = 0.91$ (refs 10, 12, 32). The average charge density, a possible measure of transiently transfected motor protein density in TSA201 cells, was $5,360 \mu\text{m}^{-2}$. The numerical values that characterize the charge movement ($V_{1/2}$, z) are quite similar to those obtained for OHCs¹⁵. The maximum charge and charge density are less in the transfected cells, even though the linear capacitance is about the same as that of an average OHC.

In OHCs, nonlinear capacitive current and electromotility can be reversibly blocked by sodium salicylate^{33–35}. As further evidence that transient currents stem from the transfected motor protein, salicylate was locally applied to cells that had been shown to present charge movement. Sodium salicylate (10 mM) reduced the transient currents to $15.5 \pm 2.9\%$ of control ($n = 8$, Fig. 5d). Of these, three cells showed recovery after 2–5 min of washing (to $88.4 \pm 8.8\%$ of control value). Finally, the existence of nonlinear capacitance was tested in 15 cells that were transfected with both GFP and human pendrin cDNA. None of the GFP-positive cells showed nonlinear capacitance. This result is significant because of the similarity of pendrin and prestin, indicating the specificity of our candidate motor protein. It also indicates that the electrophysiological results are not simply the consequence of the introduction of an anion transport protein.

Of the two manifestations of motor function—nonlinear capacitance and motility—the former is much easier to measure. Although electromotility has been shown in patches isolated from OHC basolateral membrane³⁶, the cell's cortical cytoskeleton facilitates the conversion of mechanical molecular events to gross cellular deformation^{6,37,38}. In TSA201 cells, the organized cytoskeletal network is probably missing. Consequently, even if the motor protein is expressed, gross cellular deformation is expected to be very small. Furthermore, OHCs are efficient producers of axial motility because of their cylindrical shape. Fast mechanical effects occur at constant cell volume and are generally assumed to be the result of a change in cell-surface area, owing to the aggregate conformational shape changes of large numbers of motor proteins^{6,14}. A spherical cell cannot change its surface area while its volume remains constant, whereas a cylindrical cell can³⁸. So that the generally spherical TSA201 cells might produce measurable motility, we distorted them to alter their shape by drawing them into a suction pipette (microchamber³⁹), thereby producing a dumbbell or hourglass shape (Fig. 6a). In spite of these difficulties, and our inability to test for GFP co-transfection (our motility measurement setup lacks fluorescence capability), we succeeded in demonstrating electromotility in 7 out of 31 cells tested. Two of the best responses are shown in Fig. 6b, along with a trace from a non-transfected, non-responsive cell. Our usual electrical driving signal was a brief 200-Hz sinusoid. Sodium salicylate produces a reversible decrease in the electromotile response (Fig. 6c), just as it does in OHCs^{33–35}. To examine motility at higher frequencies, comparisons between responses at 200 and 1,000 Hz are shown in Fig. 6d and e. When corrected for the characteristics of the recording system, the two responses are essentially identical (Fig. 6e), consistent with OHC data obtained in the past⁷.

Implications

We have used a logical strategy to isolate the cDNA that is responsible for the unique electromechanical function of OHCs,

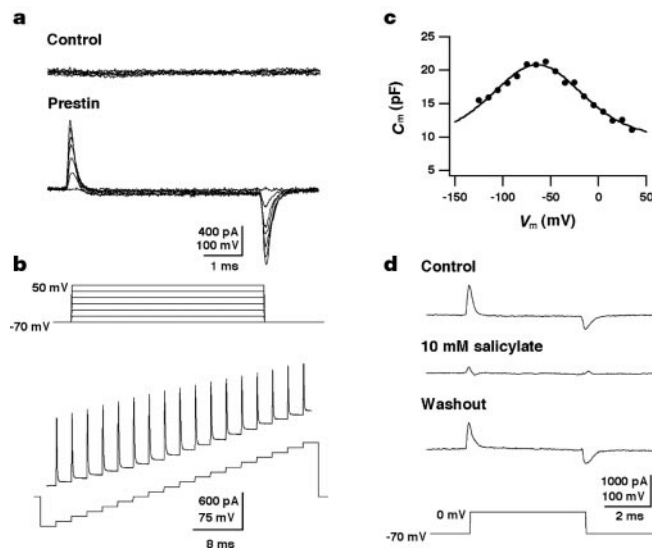


Figure 5 Voltage-dependent charge movement in TSA201 cells. Recordings were made 24–60 h after transfection. **a**, Charge movement measured as transient capacitive currents using a standard subtraction technique ($P/4$). The cell was held at -70 mV; voltage steps are 20 mV to $+50$ mV. **b**, Stair-step protocol quantifies the voltage dependence of nonlinear capacitance. Voltage was stepped from -130 to $+40$ mV in 10 mV increments of 2.5 ms each. Twenty presentations were averaged. **c**, The membrane capacitance-voltage curve was obtained from the stair-step current trace in panel **b**. Data points were fitted with the derivative of a Boltzmann function (equation (1)). **d**, Localized application of 10 mM sodium salicylate reversibly blocked the charge movement.

and have shown that prestin conveys mechanical responsiveness to a heterologous system, which is almost indistinguishable from that measured in OHCs.

The fact that a portion of the human *PRES* gene has been sequenced and mapped allows us to predict the location of the motor protein in relation to the other members of its family and to consider any relationship to genetic loci known to be related to hearing disorders or deafness. The bacterial artificial chromosome clone (RG107G13) which contains a portion of the *PRES* gene at its 3' end has a single known other gene, *RELN*, which is about 50 kb centromeric to the *PRES* gene fragment. This location is ~3.7 Mb centromeric to the location of the *PENNDIN* (*PDS*) and *DRA* genes. Thus, it appears likely that *PRES* is located in an as yet unsequenced portion of 7q31, centromeric to *PDS* and *DRA* and 50 kb telomeric to the *RELN* gene. At present there are two autosomal recessive loci (DFNB) mapped to 7q31 in addition to the *PDS* gene: DFNB14 (ref. 40) and DFNB17 (ref. 41). DFNB14 is the most likely candidate to be *PRES*, because it maps at a centromeric position to the *PDS* gene locus, whereas DFNB17 maps to the telomeric side of *PDS* and *DRA*. The proof that DFNB14 is itself distinct from *PDS* is derived from

the fact that the *PDS* exons were sequenced and appeared to be normal, indicating that the causative mutation was distinct from that seen in Pendred's syndrome⁴⁰. The clinical phenotype in this kindred is a congenital, sensorineural, autosomal recessive form of non-syndromic deafness, which is consistent with our prediction of prestin's functional role.

Our results support earlier work on OHCs, which showed that neither the subsurface cortical lattice nor the subsurface cisterns is essential for electromotility^{36–38}. It is unlikely that either of these organelles would be present in TSA201 cells, in either their native or transfected form. Our demonstration of nonlinear capacitance and motility simplifies the concept of electromechanical changes in OHCs and obviates the need for exotic schemes. Presumed conformational changes of prestin, resulting in a change of cell-surface area, produce dimensional changes in the cell. This is the basic electromotile response^{2–5}. It is yet to be determined whether stiffness changes accompany these shape changes in TSA201 cells, as they do in OHCs, thus expressing the full electromechanical process.

The reliance of the mammalian cochlea on local, OHC-based amplification is widely accepted, but there is no universal agreement

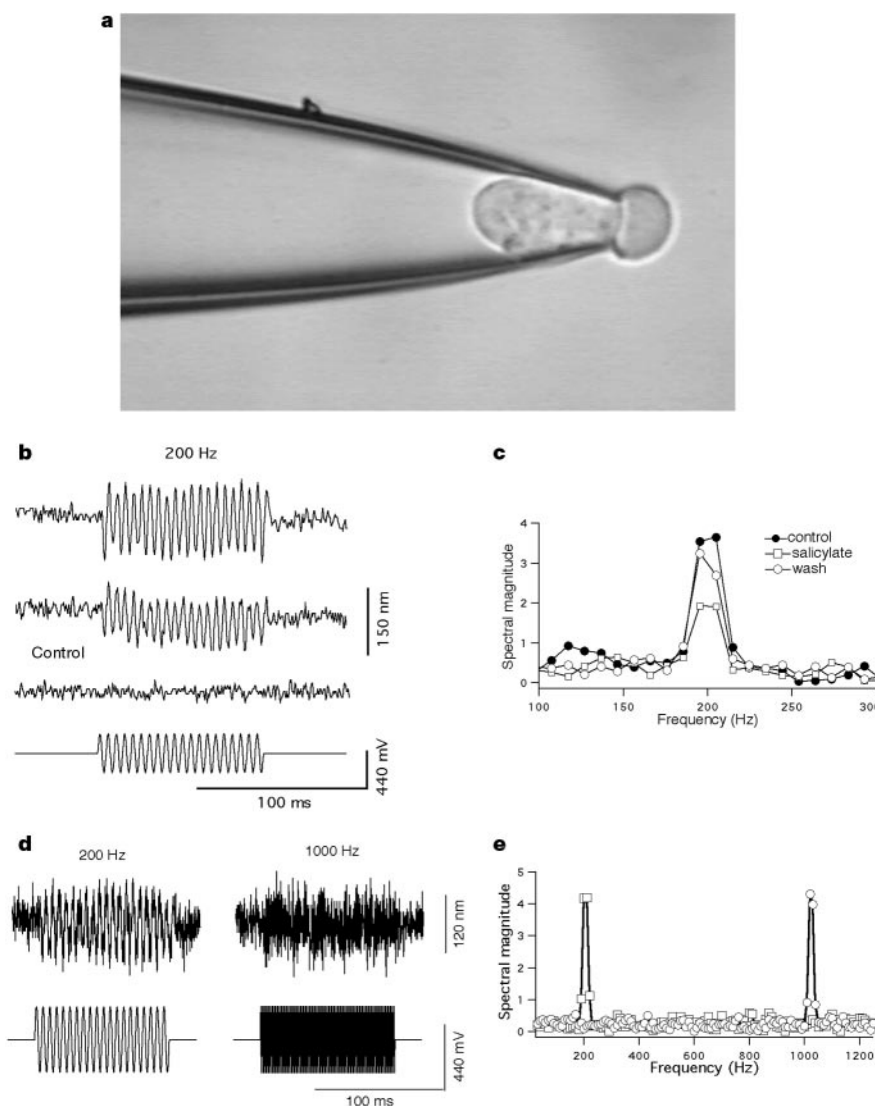


Figure 6 Examples of voltage-dependent motility expressed in TSA201 cells. **a**, Video image of a TSA201 cell partially drawn into a microchamber. Motility was measured at the farthest excluded membrane segment. **b**, Top two traces, motile responses from two transfected cells. Third trace, lack of motile response from cell transfected with control plasmid only. Fourth trace, stimulus waveform. **c**, Effect of 10 mM sodium salicylate in the

external bathing solution on the response, shown by Fourier transforms of the responses. **d**, Response (top) and stimulus (bottom) waveforms with two different command frequencies. **e**, Fourier transforms of the two response segments in **d**, including 1.2-dB correction for the system's frequency response. Response waveforms are the averages of 200 presentations. **c** and **e** are plotted with arbitrary ordinates.

about the amplifying mechanism. One view is that, powered by the cell's receptor potential, OHC electromotility provides mechanical feedback and thereby amplification^{1,12}. This view places the motor process in the cell's basolateral membrane and calls for a motor protein to drive somatic shape changes. An alternative concept is that amplification arises as a byproduct of the cell's forward transducer process and thus resides in the stereocilia⁴². Whichever mechanism dominates, the existence of electromechanical action in OHCs is indisputable. We have now identified the gene that codes for a specialized motor protein that produces this electromechanical action when expressed in cells that in their native form do not show this phenomenon.

Isolated OHCs are capable of producing an average maximum axial isometric (stall) force of ~6 nN (D.Z.Z.H. and P.D., unpublished data and ref. 43). From the number of molecules producing this force it can be estimated that the individual molecular stall force is on the order of 2.4 pN. In comparison, the stall force of kinesin is 5–6 pN (ref. 44). We note the potential of prestin to perform as a fast, voltage-driven actuator, individually or in assemblies, forming the basis of futuristic nanomachinery. □

Methods

Tissue preparation and *Pres* cDNA isolation by subtractive PCR hybridization

Animal care and use procedures were approved by the Northwestern University Institutional Review Board. Mature gerbils were killed and the organ of Corti and associated basilar membrane were isolated as described⁴⁵. After brief trypsin treatment and trituration, OHCs and IHCs were isolated in a low-calcium solution. Using a glass pipette, the cells were separated depending on their morphological appearance and were collected in LiCl buffer. mRNA was isolated from tissue using 20 µl oligo-dT magnetic beads (Dyna). cDNA pools were created with Superscript II RNAase H⁻ reverse transcriptase (Gibco) followed by amplification with a 5' cap and oligo dT-dependent PCR technique according to manufacturer's specifications (SMART PCR, Clontech). These cDNA pools were used for PCR subtractive hybridization, qualitative PCR and virtual northern (cDNA) dot-blot hybridization experiments. The PCR subtractive hybridization procedures, control reactions and subcloning were according to the manufacturer's instructions (PCR Select cDNA Subtraction Kit, Clontech).

Clones were subjected to a differential screening procedure using a PCR-Select Differential Screening Kit (Clontech). PCR-amplified cDNA inserts from these clones were denatured and vacuum filtered onto four identical nylon membranes. Each cDNA sample was separately hybridized with one of the following probes: forward subtracted (OHC–IHC) cDNA; reverse subtracted (IHC–OHC) cDNA; unsubtracted OHC cDNA; and unsubtracted IHC cDNA. These cDNA pools were random primed labelled with [³²P]dCTP and hybridized to the blots. Positively hybridizing clones were scored for differential hybridization with forward (OHC–IHC) and reverse (IHC–OHC) subtracted probes. This differential expression was confirmed by correlation with simultaneous differential hybridization with the original unsubtracted OHC and IHC cDNA pools. Sequencing was carried out using dRhodamine terminator cycle sequencing (ABI Prism) and an automated DNA sequencer (Model 377, Perkin Elmer). DNA and amino-acid sequences were compared and analysed using sequence analysis software of GCG and TMHMM²⁷ and Tmpred²⁸.

Virtual northern (cDNA) dot-blot hybridization

SMART cDNA (0.5 µg) from gerbil thyroid, IHC, OHC, and 0-, 6-, 12-, 16- and 20-day-old organs of Corti was denatured and vacuum filtered onto membranes. A *Pres* cDNA fragment (1,067-bp *Nco*I restriction fragment) was radioactively labelled with [³²P]dATP and used as a probe (Strip-EZ Kit and UltraHyb buffer, Ambion).

PCR

PCR reactions used templates of 200 ng of SMART cDNA from gerbil thyroid, newborn and adult organ of Corti, isolated IHC and organ of Corti from various postnatal ages (0 to 20 DAB). *Pres*-specific primers (sense; 5'-TACCTCACGGAGCCGCTGGT-3'; antisense; 5'-GCAGTAATCAGTCGCGTAGTCC-3') were used to amplify an 863-bp fragment from the cDNA. Cyclophilin primers (sense; 5'-TGGCAGAGGAGAAAGAGATC-3'; antisense; 5'-AAAGGGCTTCTCCACCTCGATC-3') that amplify a 301-bp DNA fragment were used as internal control. Cycle parameters were 3 min at 94 °C followed by 25–30 cycles of 94 °C for 45 s, 56 °C for 45 s and 72 °C for 1 min, with a final extension at 72 °C for 10 min.

Transient transfection

TSA201 cells, clones of human embryonic kidney 293 cells that express the simian virus 40 large T-antigen⁴⁶ in a stable manner, were cultured in D-MEM with 5% fetal calf serum, 100 U of penicillin per ml and 100 µg of streptomycin per ml. Cells were plated for 24 h before calcium phosphate transfection⁴⁷. The transfection reaction mixture contained: (1) control plasmid, pcDNA3.1; (2) 2 µg of *Pres* cDNA plasmid; (3) 2 µg *Pres* cDNA plasmid

plus 0.2 µg of GFP plasmid as transfection marker; or (4) 2 µg pendrin cDNA plasmid plus 0.2 µg GFP plasmid.

Functional verification

After 24 h the transfected cells were used for electrophysiological and physical measurements. Large, rounded, non-clustered TSA201 cells, exhibiting considerable granulation, were selected for electrophysiological and motility experiments.

Capacitance measurements

Whole-cell voltage-clamp recordings were made with an Axopatch 200B amplifier (Axon Instruments). Recording pipettes had open tip resistances of 2–3 MΩ and were filled with internal solution containing (in mM): 140 CsCl, 2 MgCl₂, 10 EGTA, 10 HEPES; pH 7.2. The external solution contained (in mM): 120 NaCl, 20 TEA-Cl, 2 CoCl₂, 2 MgCl₂, 10 HEPES, 5 glucose; pH 7.2. Osmolarity was adjusted to 300 mosm l⁻¹ with glucose. Capacitive currents were filtered at 5 kHz and digitized at 50 kHz using pClamp 7.0 software (Axon Instruments). Voltage-dependent capacitance was measured using two methods. First, a standard P/4 linear subtraction procedure¹¹ was used to detect the presence of nonlinear charge movement. Second, after nonlinear transient current was detected, a stair-step voltage protocol was used to obtain the parameters of charge movement¹⁵. For each voltage step, the measured membrane capacitance (*C_m*) was plotted as a function of membrane voltage (*V_m*) and fitted with the derivative of a Boltzmann function:

$$C_m = \frac{Q_{\max} \alpha}{\exp[\alpha(V_m - V_{1/2})](1 + \exp[-\alpha(V_m - V_{1/2})])^2} + C_{\text{lin}} \quad (1)$$

where *Q_{max}* is maximum charge transfer, *V_{1/2}* is the voltage at which the maximum charge is equally distributed across the membrane, *C_{lin}* is linear capacitance and $\alpha = ze/kT$ is the slope factor of the voltage dependence of charge transfer where *k* is Boltzmann's constant, *T* is absolute temperature, *z* is valence and *e* is electron charge. From the stair-step analysis, we have obtained the membrane resistance *R_m* = 130.45 ± 20.14 MΩ (*n* = 7).

Motility measurements

TSA201 cells were drawn into a microchamber³⁹ with gentle suction (Fig. 6a). The microchamber was mounted on a 3D micromanipulator attached to the stage of a Zeiss inverted microscope. The most distal part of the excluded segment of the cell was imaged onto a photodiode with a rectangular slit interposed into the light path. A change of photocurrent signified cell extension or contraction. The signal was current-to-voltage transformed, amplified, A/D converted and averaged (*n* = 200). The command voltage, delivered between the electrolyte solutions (L15, Gibco) surrounding and filling the microchamber, produces different voltage drops on the included and excluded cell-membrane segments, determined by the electrical voltage-divider effect of these membranes¹⁴. The voltage-divider ratio is not known; consequently we do not report actual driving voltages in the motility experiments. During motility measurements, the location of the cell within the slit was monitored by a video camera placed behind the slit. The appearance of the whole cell in the microchamber was also continuously displayed with the aid of a second video camera. This permitted independent verification of a lack of movement of the microchamber itself.

Received 6 December 1999; accepted 28 March 2000.

- Dallos, P. In *The Cochlea* (eds Dallos, P., Popper, A. N. & Fay, R. R.) 1–43 (Springer, New York, 1996).
- Brownell, W. E., Bader, C. R., Bertrand, D. & de Ribaupierre, Y. Evoked mechanical responses of isolated outer hair cells. *Science* **227**, 194–196 (1985).
- Ashmore, J. F. A fast motile response in guinea-pig outer hair cells: the cellular basis of the cochlear amplifier. *J. Physiol. (Lond.)* **388**, 323–347 (1987).
- Santos-Sacchi, J. & Dilger, J. P. Whole cell currents and mechanical responses of isolated outer hair cells. *Hearing Res.* **35**, 143–150 (1988).
- Kachar, B., Brownell, W. E., Altschuler, R. A. & Fex, J. Electrokinetic shape changes of cochlear outer hair cells. *Nature* **322**, 365–368 (1986).
- Holley, M. C. & Ashmore, J. F. On the mechanisms of a high frequency force generator in outer hair cells isolated from the guinea pig cochlea. *Proc. R. Soc. Lond. Ser. B* **232**, 413–429 (1988).
- Dallos, P. & Evans, B. N. High frequency motility of outer hair cells and the cochlear amplifier. *Science* **267**, 2006–2009 (1995).
- Frank, G., Hemmert, W. & Gummer, A. W. Limiting dynamics of high-frequency electromechanical transduction of outer hair cells. *Proc. Natl Acad. Sci. USA* **96**, 4420–4425 (1999).
- Ashmore, J. F. In *Mechanics of Hearing* (eds Kemp, D. & Wilson, J. P.) 107–113 (Plenum, New York, 1999).
- Santos-Sacchi, J. Reversible inhibition of voltage-dependent outer hair cell motility and capacitance. *J. Neurosci.* **11**, 3096–3110 (1991).
- Armstrong, C. M. & Bezanilla, F. Charge movement associated with the opening and closing of the activation gates of the Na channels. *J. Gen. Physiol.* **63**, 533–552 (1974).
- Ashmore, J. F. In *Sensory Transduction* (eds Corey, D. P. & Roper, S. D.) 395–412 (Rockefeller Univ. Press, New York, 1992).
- He, D. Z. Z. & Dallos, P. Somatic stiffness of cochlear outer hair cells is voltage dependent. *Proc. Natl Acad. Sci. USA* **96**, 8223–8228 (1999).
- Dallos, P., Evans, B. N. & Hallworth, R. On the nature of the motor element in cochlear outer hair cells. *Nature* **350**, 155–157 (1991).
- Huang, G. & Santos-Sacchi, J. Mapping the distribution of the outer hair cell motility voltage sensor by electrical amputation. *Biophys. J.* **65**, 2228–2236 (1993).
- Forge, A. Structural features of the lateral walls in mammalian cochlear outer hair cells. *Cell Tissue Res.* **265**, 473–483 (1991).

17. Kalinec, F. & Kachar, B. Inhibition of outer hair cell electromotility by sulphhydryl specific reagents. *Neurosci. Lett.* **157**, 231–234 (1993).
18. Knipper, M. *et al.* Immunological identification of candidate proteins involved in regulating shape changes of outer hair cells. *Hearing Res.* **86**, 100–1100 (1995).
19. Géléoc, G. S. G., Casalotti, S. O., Forge, A. & Ashmore, J. F. A sugar transporter as a candidate for the outer hair cell motor. *Nature Neurosci.* **2**, 713–719 (1999).
20. He, D. Z. Z., Evans, B. N. & Dallos, P. First appearance and development of electromotility in neonatal gerbil outer hair cells. *Hearing Res.* **78**, 77–90 (1994).
21. Diatchanko, L. *et al.* Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc. Natl Acad. Sci. USA* **93**, 6025–6030 (1996).
22. Kozak, M. An analysis of 5' noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15**, 8125–8148 (1987).
23. Everett, L. A. *et al.* Pendred syndrome is caused by mutations in a putative sulphate transporter gene (PDS). *Nature Genet.* **17**, 411–422 (1997).
24. Moseley, R. H. *et al.* Downregulated in adenoma gene encodes a chloride transporter defective in congenital chloride diarrhea. *Am. J. Physiol.* **276**, 185–192 (1999).
25. Scott, D. A., Wang, R., Kreman, T. M., Sheffield, V. C. & Karniski, L. P. The Pendred syndrome gene encodes a chloride–iodide transport protein. *Nature Genet.* **4**, 440–443 (1999).
26. Everett, L. A. & Green, E. D. A family of mammalian anion transporters and their involvement in human genetic diseases. *Hum. Mol. Genet.* **8**, 1883–1891 (1999).
27. Sonnenhammer, E. L. L., Heijne, G. von & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. 6th Int. conf. Intelligent Syst. Mol. Biol. (ISMB98)* (eds Glasgow, J. *et al.*) 175–182 (Amer. Assoc. for Artificial Intelligence Press, Menlo Park, CA, 1998).
28. Hofman, K. & Stoffel, W. Tmbase—a database of membrane spanning protein segments. *Biol. Chem.* **374**, 166 (1993).
29. Sandal, N. N. & Marcker, K. A. Similarities between a soybean nodulin, *Neurospora crassa* sulphate permease II and a putative human tumour suppressor. *Trends Biochem. Sci.* **19**, 19 (1994).
30. Holley, M. C., Kalinec, F. & Kachar, B. Structure of the cortical cytoskeleton in mammalian outer hair cells. *J. Cell. Sci.* **102**, 569–580 (1992).
31. Souter, M., Nevill, G. & Forge, A. Postnatal development of membrane specialisations of gerbil outer hair cells. *Hearing Res.* **91**, 43–62 (1995).
32. Iwasa, K. H. Effect of stress on the membrane capacitance of the auditory outer hair cell. *Biophys. J.* **65**, 492–498 (1993).
33. Shehata, W., Brownell, W. E. & Dieler, R. Effects of salicylate on shape, electromotility and membrane characteristics of isolated hair cells from the guinea pig cochlea. *Acta Oto-Laryngol. (Stockholm)* **111**, 707–718 (1991).
34. Kakehata, S. & Santos-Sacchi, J. Effects of salicylate and lanthanides on outer hair cell motility and associated gating charge. *J. Neurosci.* **16**, 4881–4889 (1996).
35. Tunstall, M. J., Gale, L. E. & Ashmore, J. F. Action of salicylate on membrane capacitance of outer hair cells from the guinea-pig cochlea. *J. Physiol. (London)* **485**, 739–752 (1995).
36. Kalinec, F., Holley, M. C., Iwasa, K., Lim, D. J. & Kachar, B. A membrane-based force generation mechanism in auditory sensory cells. *Proc. Natl Acad. Sci. USA* **89**, 8671–8675 (1992).
37. Huang, G. J. & Santos-Sacchi, J. Motility voltage sensor of the outer hair cell resides within the lateral plasma membrane. *Proc. Natl Acad. Sci. USA* **91**, 12268–12272 (1994).
38. Adachi, M. & Iwasa, K. H. Electrically driven motor in the outer hair cell: effect of a mechanical constraint. *Proc. Natl Acad. Sci. USA* **96**, 7244–7249 (1999).
39. Evans, B. N., Dallos, P. & Hallworth, R. in *Cochlear Mechanisms* (eds Wilson, J. P. & Kemp, D. T.) 205–206 (Plenum, London, 1989).
40. Mustapha, M. *et al.* Identification of a locus on chromosome 7q31, DFNB14, responsible for prelingual sensorineural non-syndromic deafness. *Eur. J. Hum. Genet.* **6**, 548–551 (1998).
41. Greinwald, J. H. Jr *et al.* Localization of a novel gene for nonsyndromic hearing loss (DFNB17) to chromosome region 7q31. *Am. J. Med. Genet.* **78**, 107–113 (1998).
42. Hudspeth, A. J. Mechanical amplification of stimuli by hair cells. *Curr. Opin. Neurobiol.* **7**, 480–486 (1997).
43. Iwasa, K. H. & Adachi, M. Force generation in the outer hair cell of the cochlea. *Biophys. J.* **73**, 546–555 (1997).
44. Svoboda, K. & Block, S. M. Force and velocity measured for single kinesin molecules. *Cell* **77**, 773–784 (1994).
45. He, D. Z. Z. Relationship between the development of outer hair cell electromotility and efferent innervation: a study in cultured organ of Corti of neonatal gerbils. *J. Neurosci.* **15**, 3634–3643 (1997).
46. Margolskee, R. F., McHendry-Rinde, B. & Horn, R. Panning transfected cells for electrophysiological studies. *Biotechniques* **15**, 906–911 (1993).
47. Graham, F. L. & van der Eb, A. J. Transformation of rat cells by DNA of human adenovirus 5. *Virology* **54**, 536–539 (1973).

Acknowledgements

We thank L. H. Pinto and J. S. Takahashi for their comments on the manuscript. We also thank P. Kopp for providing the human pendrin cDNA, B. Schulte for providing the λ gt11 gerbil cochlea library, and Y. Tang, M. Brenner and J. Cheng for technical assistance. This work was primarily supported by a Senior Fellowship from the McKnight Endowment Fund for Neuroscience to P.D. and the National Institute on Deafness and other Communication Disorders, NIH.

Correspondence and requests for materials should be addressed to P.D. (e-mail: p-dallos@nwu.edu). Pres has been deposited in the GenBank under accession number AF230376.

Cosmic γ -ray background from structure formation in the intergalactic medium

Abraham Loeb* & Eli Waxman†

* Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

† Department of Condensed Matter Physics, Weizmann Institute, Rehovot, 76100, Israel

The Universe is filled with a diffuse background of γ -ray radiation¹, the origin of which remains one of the unsolved puzzles of cosmology. Less than one-quarter of the γ -ray flux can be attributed to unresolved discrete sources^{2,3}, such as active galactic nuclei; the remainder appears to constitute a truly diffuse background. Here we show that the shock waves induced by gravity in the gas of the intergalactic medium, during the formation of large-scale structures like filaments and sheets of galaxies, produce a population of highly relativistic electrons. These electrons scatter a small fraction of the cosmic microwave background photons in the local Universe up to γ -ray energies, thereby providing the γ -ray background. The predicted diffuse flux agrees with the observed background across more than four orders of magnitude in photon energy, and the model predicts that the γ -ray background, though generated locally, is isotropic to better than five per cent on angular scales larger than a degree. Moreover, the agreement between the predicted and observed background fluxes implies a mean cosmological density of baryons that is consistent with Big Bang nucleosynthesis.

The Universe started from a smooth initial state, with small density fluctuations that grew in time due to the effect of gravity. The formation of structure resulted in non-relativistic, collisionless shock waves in the baryonic gas due to converging bulk flows, and raised its mean temperature to $\approx 10^7$ K (≈ 1 keV) at the present time⁴. This is indeed the characteristic temperature of the warm gas in groups of galaxies, which are the typical objects forming at the present epoch. The intergalactic shocks are usually strong, because the gravitationally induced bulk flows are often characterized by a high Mach number. The intergalactic gas may have also been heated by shocks driven by outflows from young galaxies^{5–7}.

Collisionless, non-relativistic, shocks are known to generically accelerate a power-law distribution of relativistic electrons, with a number density, n_e , per electron momentum, p_e , of $dn_e/dp_e = Kp_e^{-\alpha}$, where K is a constant⁸. The power-law index for strong shocks in a gas with an adiabatic index of $\Gamma = 5/3$, is $\alpha = [(r+2)/(r-1)] = 2$ (refs 8–10), where $r = [(\Gamma+1)/(\Gamma-1)]$ is the shock compression factor. Such an electron distribution is found in the strong shocks surrounding supernova remnants in the interstellar medium⁸. Recent X-ray^{11,12} and TeV (refs 13, 14) observations of the supernova remnants SN1006 and SNR RX J1713.7–3946 imply that electrons are accelerated in the remnant shocks up to an energy ≈ 100 TeV and are confined to the collisionless fluid by magnetic fields. These shocks have a velocity of the order of 10^3 km s⁻¹, similar to the velocity of the intergalactic shocks we consider here. The inferred energy density in relativistic electrons constitutes 1–10% of the post-shock energy density in these remnants, a fraction consistent with the global ratio between the mean energy density of cosmic-ray electrons and the turbulent energy density in the interstellar medium of our galaxy.

As the physical processes of shock acceleration can be scaled up to intergalactic distances, it appears natural to assume that a similar population of relativistic electrons was also produced in the intergalactic medium. The maximum Lorentz factor of the accelerated

electrons, $\gamma_{\max}$, is set by equating their acceleration and cooling times. The e-folding time for the acceleration of the relativistic electrons is $t_{\text{acc}} \approx (r_L c/v_{\text{sh}}^2)$, where $v_{\text{sh}} = (\Gamma+1)[kT/(\Gamma-1)m_p]^{1/2}$ is the shock velocity relative to the unshocked gas, m_p is the proton mass, and r_L is the electron Larmor gyration radius. For an electron with a Lorentz factor γ_e , $r_L = 5.5 \times 10^{-2} (\gamma_e/B_{-7})$ pc, where $\gamma_e = (\gamma_e/10^7)$ and B_{-7} is the magnetic field strength in units of 0.1 μ G. A magnetic field amplitude of 0.1–1 μ G is often detected in the haloes of galaxy clusters^{15–17}. In particular, a magnetic field amplitude of ≥ 0.1 μ G was inferred on the multi-megaparsec scale of the Coma-A1367 supercluster¹⁸, which provides a good example for the intergalactic structures of interest here. For such magnetic field amplitude we get $t_{\text{acc}} \approx 2 \times 10^4 \text{ yr } (\gamma_e/B_{-7})(kT/\text{keV})^{-1}$.

The acceleration time is much shorter than the lifetime of the intergalactic shocks, which is comparable to the age of the Universe. The maximum Lorentz factor of the electrons, $\gamma_{\max}$, is therefore not limited by the lifetime of their accelerator, but rather by their cooling, primarily due to Compton scattering off the cosmic microwave background¹⁹. Synchrotron cooling is negligible for $B_{-7} \lesssim 10$. The characteristic cooling time due to inverse-Compton scattering is $t_{\text{cool}} = [m_e c/(4/3)\sigma_T \gamma_e u_{\text{cmb}}] = 1.2 \times 10^{10} \text{ yr } (\gamma_e/200)^{-1}$, where m_e is the electron mass, σ_T is the Thomson cross-section, and u_{cmb} is the energy density of the cosmic microwave background. By equating the acceleration and cooling times, $t_{\text{acc}} = t_{\text{cool}}$, we find $\gamma_{\max} = 3.7 \times 10^7 [B_{-7}(kT/\text{keV})]^{1/2}$.

The estimates given above are valid as long as the electron Larmor radius is smaller than the coherence length of the magnetic field. Because the Larmor radius of the relativistic electrons is extremely

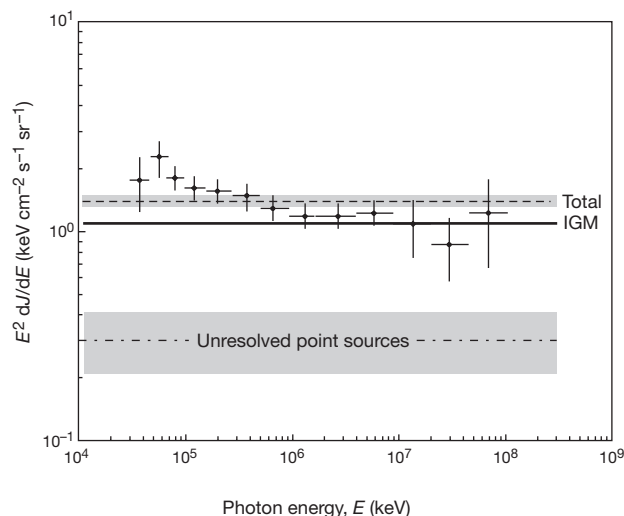


Figure 1 Spectrum of the unresolved γ -ray background. Points with error bars are the observed data from EGRET¹. The lower shaded region shows the expected contribution from unresolved point sources, based on empirical modelling of the luminosity function of blazars^{2,3}. The solid line shows the diffuse emission from shocks in the intergalactic medium (IGM) according to equation (1) with $\xi_e = 0.05$ and $f_{\text{sh}} kT = 1$ keV. The upper shaded region shows the sum of these contributions, which provides an excellent fit to the data. The blazar contribution was calibrated^{2,3} only by the total flux in photons with energies $E > 10^5$ keV; hence, the small deviations of some data points from the upper shaded region might be due to uncertainties in the cumulative blazar spectrum. Although the background flux predicted by equation (1) is independent of the magnetic field strength in the intergalactic shocks, the maximum energy of scattered photons does depend on the field strength and is given by $h\nu_{\max} \approx 1.2 B_{-7} (kT/\text{keV})$ TeV. Photons with energies ≥ 1 TeV produce an electron–positron pair as they scatter on the infrared background^{28,29}. The pair cools again by scattering microwave photons, which may in turn produce new pairs. The energy originally stored in photons with $h\nu \geq 1$ TeV is spread smoothly over the 100 MeV to 1 TeV energy range³⁰, and might raise the existing flux there. However, we expect this effect to be small, because in our model only a small fraction of the electrons are accelerated to the energy required for the production of $h\nu \geq 1$ TeV photons.

small, $r_L \approx 0.1$ pc, this condition is satisfied even if the field coherence length is much shorter than the megaparsec scale of the intergalactic shocks.

All electrons with $\gamma_e > 200$ lose their energy over the age of the Universe. The energy of these electrons is converted to a diffuse background of photons, which are produced through inverse Compton scattering of microwave background photons. The initial energy of a microwave photon, $h\nu_0$, is boosted by the scattering up to an average value of $h\nu = (4/3)\gamma_e^2 h\nu_0$. Substituting the mean frequency of the microwave background photons for ν_0 , we find that the accelerated intergalactic electrons produce a diffuse background of radiation extending from the ultraviolet ($h\nu = 36(\gamma_e/200)^2$ eV) up to extreme γ -ray energies ($h\nu = 89\gamma_e^2$ GeV). For a power-law distribution of relativistic electrons, $dn_e/dp_e = Kp_e^{-2}$, the energy density of the scattered radiation is predicted to be constant per logarithmic frequency intervals, $\nu(d\nu/d\nu) = Kc/2 = u_e/2 \ln \gamma_{\max}$. Here $u_e = Kc \ln \gamma_{\max}$ is the energy density in relativistic electrons produced by the intergalactic shocks. If a fraction, f_{sh} , of the baryons in the Universe were shocked to a (mass-weighted) mean temperature T , and a fraction, ξ_e , of the shock thermal energy was transferred to relativistic electrons, then $u_e = \frac{3}{2} \xi_e f_{\text{sh}} n_p kT = 5.1 \times 10^{-16} \xi_e f_{\text{sh}} (\Omega_b h_{70}^2 / 0.04) (kT/\text{keV}) \text{ erg cm}^{-3}$, where n_p is the average proton density, Ω_b is the cosmological baryon density parameter, and h_{70} is the Hubble constant in units of $70 \text{ km s}^{-1} \text{ Mpc}^{-1}$. By substituting $\ln \gamma_{\max} = \ln(4 \times 10^7)$, we then get

$$E^2 \frac{dJ}{dE} = 1.1 \left(\frac{\xi_e}{0.05} \right) \left(\frac{\Omega_b h_{70}^2}{0.04} \right) \left(\frac{f_{\text{sh}} kT}{\text{keV}} \right) \text{keV cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1} \quad (1)$$

where $E = h\nu$, and J is the number flux of photons per solid angle. Adopting the result from hydrodynamic simulations of $f_{\text{sh}}(kT/\text{keV}) \sim 1$ (ref. 4), the Big Bang nucleosynthesis value of $\Omega_b h_{70}^2 = 0.04 \pm_{0.027}^{0.014}$ (refs 20, 21; based on the deuterium abundance), and the value of $\xi_e \approx 0.05$ inferred from non-relativistic collisionless shocks in the interstellar medium, we obtain a background flux of $\sim 1 \text{ keV cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1}$ in the photon energy range of 3 MeV to 100 GeV. As Fig. 1 illustrates, this result is in excellent agreement with the γ -ray background detected by the EGRET instrument on board the CGRO satellite¹.

The sky-averaged contribution of identified extragalactic sources, such as blazars (γ -ray-loud active galactic nuclei), amounts to only $\lesssim 7\%$ of the extragalactic γ -ray flux³. As shown in Fig. 1, the recent EGRET data is consistent with a luminosity function of faint, undetected blazars, which could account for only $\lesssim 25\%$ of the unresolved γ -ray background²³. The remainder of the hard γ -ray background is expected to be highly isotropic in our model because it is integrated over a large volume throughout the Universe. Its large-angle fluctuations should be comparable to that of the hard X-ray background, of which $\sim 40\%$ is contributed by active galaxies at $z \lesssim 1$ (ref. 22), because these galaxies trace the same large-scale structure as painted by the cosmic web of intergalactic shocks. This implies a root-mean-square fluctuation amplitude smaller than $\sim 5\%$ on angular scales larger than a degree²³. The predicted high level of isotropy can serve as a test of our model, as soon as the isotropy of the γ -ray background is measured to a higher precision than at present available. Another possible test is the spectral distortion induced in the microwave background band by the low-energy tail of the non-thermal intergalactic electrons. This distortion is, however, well below the upper limit inferred from the COBE satellite data²⁴.

At photon energies $\lesssim 1$ MeV the scattered background is weaker than the cumulative flux produced by discrete sources, such as active galactic nuclei. In particular, the predicted scattered flux amounts only to ~ 10 – 15% of the soft X-ray background measured by ROSAT at 1–2 keV, and is consistent with the upper limit of ~ 20 – 30% on its unresolved fraction^{22,25}.

Most of the γ -ray background is produced during the latest episode of shock heating of the intergalactic gas at $z \lesssim 1$, as the mean gas temperature is expected to decrease rapidly with increasing redshift⁴. Furthermore, the photons emitted at redshifts $z \gtrsim 0.5$ lose a significant fraction of their energy due to the expansion of the Universe. A more accurate calculation of the γ -ray background can be achieved by simulating the hydrodynamic evolution of the intergalactic medium and injecting a population of accelerated electrons into each fluid element which passes through a shock, at all redshifts. Such a simulation can yield both the background flux and its statistical fluctuations on the sky. For a given cosmological model of structure formation, the derived background flux will be proportional to ξ_e , and will not depend on any other free parameters.

Although the warm ($\lesssim 10^7$ K) intergalactic medium is expected theoretically to include a substantial fraction of the baryons in the present-day Universe, it has not yet been observed directly. The soft X-ray emission from intergalactic structures, such as sheets and filaments or galaxy groups, is often too faint to be detectable with current telescopes⁴. The observed cosmological density of stars²¹, $\Omega_* h_{70} = 3.5 \times 10^{-3}$, is an order of magnitude smaller than the baryon density predicted by Big Bang nucleosynthesis. If the γ -ray background indeed results from ten-million-degree shocks in the intergalactic medium, then we can derive a lower limit on the mean baryon density required to produce its observed flux. This minimum is obtained by substituting $\xi_e = 1/2$ (equipartition) and $f_{\text{sh}} = 1$ in equation (1), yielding $\Omega_b h_{70}^2 \gtrsim 4 \times 10^{-3}$. The inferred lower limit is larger than the observed mass density in stars. Thus, if the reality of our model is proven, then this would not only explain the origin of the diffuse component of the γ -ray background, but also provide the first evidence for the ‘missing baryons’ (note that even if the truly diffuse component amounts to only $\sim 20\%$ of the γ -ray background flux, the more plausible range of $\xi_e \lesssim 10\%$ yields a lower limit on the baryon density which is still larger than the mass density in stars).

In our model, the γ -ray background is produced in the dense filaments and sheets which channel gas from converging bulk flows in the intergalactic medium. The densest and hottest shocks occur at the intersections of these filaments around the locations of clusters of galaxies. Although the richest accreting clusters may be rare and contribute a small fraction of the background, they contain the brightest shocks in the sky and produce the strongest fluctuations in the diffuse γ -ray background. Direct detection of these shocks could be used to calibrate ξ_e in our model. Even a statistical detection of a cross-correlation signal between background fluctuations and other sources that trace the same large-scale structure—such as galaxies, X-ray gas, the Sunyaev–Zel’dovich effect, or synchrotron emission from the high-energy electrons in the intergalactic medium—could be used to prove the validity of our model.

We predict that as cold gas goes through the strong virialization shock of an accreting cluster, it emits non-thermal radiation with photon energies between 30 eV and 1 TeV. The shocked gas loses a fraction $\sim (4.5/7) \times \xi_e \sim 3\%$ of its thermal energy in this process. For $h\nu \gg 10$ keV, the cooling time of the corresponding relativistic electrons is shorter than a billion years. In this regime the total non-thermal luminosity of the cluster, L_{nt} , is limited by the time it takes the cluster gas to cross the virialization shock, that is, $t_{\text{vir}} \approx (2 \text{ Mpc}/2 \times 10^3 \text{ km s}^{-1}) = 10^9 \text{ yr}$. For a young cluster which shocks gas of total mass M_{gas} to a virial temperature T_{gas} , the total (bolometric) non-thermal luminosity is

$$L_{\text{nt}} \approx \left(\frac{4.5\xi_e/7}{t_{\text{vir}}} \right) \left(\frac{M_{\text{gas}}}{m_p} \right) \left(\frac{3}{2} kT_{\text{gas}} \right) \\ = 1.5 \times 10^{45} \left(\frac{\xi_e}{0.05} \right) \left(\frac{10^9 \text{ yr}}{t_{\text{vir}}} \right) \left(\frac{M_{\text{gas}}}{10^{14} M_{\odot}} \right) \left(\frac{KT_{\text{gas}}}{5 \text{ keV}} \right) \text{ erg s}^{-1} \quad (2)$$

where $M_{\odot}$ is the solar mass. (The total cluster mass, including the

dark matter, is typically an order of magnitude larger than M_{gas}). The luminosity per each logarithmic interval in photon energy, anywhere between 30 eV and 1 TeV is $\sim 0.1 L_{\text{nt}}$. This emission component would affect estimates of the gas temperature and the thermal luminosity of young clusters which form through shocking of cold gas. However, the high-energy emission would be suppressed in the weak shocks that result from mergers of pre-existing clusters which contain pre-heated gas, because the energy distribution of the accelerated electrons is steeper for weak shocks⁸. As a cluster relaxes to hydrostatic equilibrium and ages, its non-thermal luminosity declines. The emission of hard radiation disappears almost entirely as soon as there is no strong shocking of gas.

At a perfectly quiescent state of hydrostatic equilibrium, the non-thermal radiation spectrum cuts-off above a photon energy, $h\nu \approx 0.5 \text{ keV} (\tau/3 \times 10^9 \text{ yr})^{-2}$, for which the electron cooling time equals the cluster age, τ . Detection of this cut-off could be used as a method of dating the time that has passed since the last significant accretion episode of cold gas by the cluster. Most relaxed clusters with ages $\tau \gg 10^9 \text{ yr}$ are expected to show only weak non-thermal emission at $h\nu \gtrsim 10 \text{ keV}$. We note that the detection of non-thermal X-ray emission has been reported recently for several clusters^{16,17,26}. Also, some X-ray clusters possess radio haloes²⁷ which may result from synchrotron emission by the residual population of relativistic electrons in the cluster magnetic field. $\square$

Received 21 January; accepted 29 March 2000.

1. Sreekumar, P. *et al.* EGRET observations of the extragalactic gamma-ray emission. *Astrophys. J.* **494**, 523–534 (1998).
2. Chiang, J. & Mukherjee, R. The luminosity function of the EGRET gamma-ray blazars. *Astrophys. J.* **496**, 752–760 (1998).
3. Mukherjee, R. & Chiang, J. EGRET γ -ray blazars: luminosity function and contribution to the extragalactic γ -ray background. *Astrophys. J.* **511**, 213–215 (1999).
4. Cen, R. & Ostriker, J. P. Where are the baryons? *Astrophys. J.* **514**, 1–6 (1999).
5. Metzler, C. A. & Evrard, A. E. A simulation of the intracluster medium with feedback from cluster galaxies. *Astrophys. J.* **437**, 564–583 (1994).
6. Nevalainen, J., Markevitch, M. & Forman, W. R. The cluster M-T relation from temperature profiles observed with ASCA and ROSAT. *Astrophys. J.* (in the press); also as preprint astro-ph/9911369 at (<http://xxx.lanl.gov>) (1999); cited 18 Nov. 1999).
7. Loewenstein, M. Heating of intergalactic gas and cluster scaling relations. *Astrophys. J.* (in the press); also as preprint astro-ph/9910276 at (<http://xxx.lanl.gov>) (1999); cited 14 Oct. 1999).
8. Blandford, R. & Eichler, D. Particle acceleration at astrophysical shocks—a theory of cosmic-ray origin. *Phys. Rep.* **154**, 1–75 (1987).
9. Bell, A. R. The acceleration of cosmic rays in shock fronts. II. *Mon. Not. R. Astron. Soc.* **182**, 147–156 (1978).
10. Blandford, R. D. & Ostriker, J. P. Particle acceleration by astrophysical shocks. *Astrophys. J.* **221**, L29–L32 (1978).
11. Koyama, K. *et al.* Evidence for shock acceleration of high-energy electrons in the supernova remnant SN1006. *Nature* **378**, 255–258 (1995).
12. Koyama, K. *et al.* Discovery of non-thermal x-rays from the northwest shell of the new SNR RX J1713.7–3946: The second SN 1006? *Publ. Astron. Soc. Jpn* **49**, L7–L11 (1997).
13. Tanimori, T. *et al.* Discovery of TeV gamma rays from SN 1006: further evidence for the supernova remnant origin of cosmic rays. *Astrophys. J.* **497**, L25–L28 (1998).
14. Muraishi, H. *et al.* Evidence for TeV gamma-ray emission from the shell type SNR RXJ1713.7–3946. *Astron. Astrophys.* (in the press); also as preprint astro-ph/0001047 at (<http://xxx.lanl.gov>) (2000); cited 5 Jan. 2000).
15. Kronberg, P. Extragalactic magnetic fields. *Rep. Prog. Phys.* **57**, 325–382 (1994).
16. Fusco-Femiano, R. *et al.* Hard x-ray radiation in the Coma cluster spectrum. *Astrophys. J.* **513**, L21–L24 (1999).
17. Rephaeli, Y., Gruber, D. & Blanco, P. Rossi X-Ray Timing Explorer observations of the Coma cluster. *Astrophys. J.* **511**, L21–L24 (1999).
18. Kim, K. T., Kronberg, P. P., Giovannini, G. & Venturi, T. Discovery of intergalactic radio emission in the Coma-A1367 supercluster. *Nature* **341**, 720–723 (1989).
19. Felten, J. E. & Morrison, P. Omnidirectional inverse Compton and synchrotron radiation from cosmic distribution of fast electrons and thermal photons. *Astrophys. J.* **146**, 686–708 (1966).
20. Tytler, D., O'Meara, J. M., Suzuki, N. & Lubin, D. Review of Big Bang nucleosynthesis and primordial abundances. *Phys. Scripta* (in the press); also as preprint astro-ph/0001318 at (<http://xxx.lanl.gov>) (2000); cited 18 Jan. 2000).
21. Fukugita, M., Hogan, C. J. & Peebles, P. J. E. The cosmic baryon budget. *Astrophys. J.* **503**, 518–530 (1998).
22. Mushotzky, R. F., Cowie, L. L., Barger, A. J. & Arnaud, K. A. Resolving the extragalactic hard X-ray background. *Nature* (in the press); also as preprint astro-ph/0002313 at (<http://xxx.lanl.gov>) (2000); cited 16 Feb. 2000).
23. Fabian, A. C. & Barcons, X. The origin of the X-ray background. *Annu. Rev. Astron. Astrophys.* **30**, 429–456 (1992).
24. Fixsen, D. J. *et al.* The cosmic microwave background spectrum from the full COBE FIRAS data set. *Astrophys. J.* **473**, 576–587 (1996).
25. Hasinger, G. *et al.* The ROSAT deep survey. I. X-ray sources in the Lockman field. *Astron. Astrophys.* **329**, 482–494 (1998).

26. Kaastra, J. S. *et al.* High and low energy nonthermal x-ray emission from the Abell 2199 cluster of galaxies. *Astrophys. J.* **519**, L119–L122 (1999).
27. Deiss, B. M., Reich, W., Lesch, H. & Wielebinski, R. The large-scale structure of the diffuse radio halo of the Coma cluster at 1.4 GHz. *Astron. Astrophys.* **321**, 55–63 (1997).
28. Primack, J. R., Bullock, J. S., Somerville, R. S. & McMinn, D. Probing galaxy formation with TeV gamma ray absorption. *Astrophys. J.* **11**, 93–102 (1999).
29. Konopelko, A. K., Kirk, J. G., Stecker, F. W. & Mastichiadis, A. Evidence for intergalactic absorption in the TeV gamma-ray spectrum of Markarian 501. *Astrophys. J.* **518**, L13–L15 (1999).
30. Coppi, B. S. & Aharonian, F. A. Constraints on the very high energy emissivity of the universe from the diffuse GeV gamma-ray background. *Astrophys. J.* **487**, L9–L12 (1997).

Acknowledgements

This work was supported in part by the Israel-US BSF and by NSF. A.L. thanks the Weizmann Institute for hospitality during the work. E.W. is the incumbent of the Beracha foundation career development chair.

Correspondence and requests for materials should be addressed to A.L. (e-mail: aloeb@cfa.harvard.edu).

Proximate humid and dry regions in Jupiter's atmosphere indicate complex local meteorology

M. Roos-Serote*, A. R. Vasavada†, L. Kamp‡, P. Drossart§, P. Irwin||, C. Nixon|| & R. W. Carlson‡

* Observatório Astronómico de Lisboa, Tapada da Ajuda, 1349-018 Lisbon, Portugal

† Department of Earth and Space Sciences, University of California at Los Angeles, Los Angeles, California 90095, USA

‡ Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California 91109, USA

§ Département Spatial (CNRS-UMR8632), Observatoire de Paris-Meudon, 5, Place Janssen, 92915 Meudon Cedex, France

|| Atmospheric Physics Department, Clarendon Laboratory, Oxford University, Parks Road, Oxford OX13PU, United Kingdom

Models of Jupiter's formation and structure predict that its atmosphere is enriched in oxygen, relative to the Sun, and that consequently water clouds should be present globally near the 5-bar pressure level^{1,2}. Past attempts to confirm these predictions have led to contradictory results^{3–5}; in particular, the Galileo probe revealed a very dry atmosphere at the entry site, with no significant clouds at depths exceeding the 2-bar level^{6,7}. Although the entry site was known to be relatively cloud-free, the contrast between the observed local dryness and the expected global wetness was surprising. Here we analyse near-infrared (around 5 μm) observations of Jupiter, a spectral region that can reveal the water vapour abundance and vertical cloud structure in the troposphere⁸. We find that humid and extremely dry regions exist in close proximity, and that some humid regions are spatially correlated with bright convective clouds extending from the deep water clouds to the visible atmosphere.

During Galileo's fourth orbit (E4) in December 1996, the Near Infrared Mapping Spectrometer (NIMS) and the Solid State Imager (SSI) on the Galileo spacecraft acquired data over the same region of Jupiter's atmosphere within a time span of 64 h. The observed region contained a '5- μm hot spot', a relatively cloud-free area where thermal radiation from the deep atmosphere escapes to space, and an evolving, bright cloud northward of it (Fig. 1). The NIMS measurements at 5 μm sound to deeper levels (~ 6 bar) than the SSI observations (~ 3 bar). The low spectral resolving power (200) of

the NIMS instrument and the weakness of the 5- μm water absorption bands preclude the direct retrieval of the deep-water abundance; however, the integrated column density of water between about 3 and 8 bar can be retrieved. The depth of the water band at 5.025 μm , D , defined as the difference between the radiance at the

continuum (out-of-band) wavelength (5.052 μm) and the centre of the water band (5.025 μm) divided by the continuum radiance, is a proxy for the amount of water vapour in this part of the troposphere, modelled as the relative humidity, RH, which is actual normalized to the saturated water vapour pressure at the ambient temperature. Model calculations indicate that D varies between about 0.23 for dry spectra ($\text{RH} \approx 1\%$) and 0.4 for very humid spectra ($\text{RH} > \sim 20\%$).

To assess the RH and cloud structure within and surrounding the hot spot, we modelled NIMS spectra using a line-by-line radiative transfer code. The cloud scattering was simplified by replacing cloud layers with absorbing/reflecting layers⁹. The water vapour vertical profiles follow the chosen RH curves, until the level where the mixing ratio ($[\text{H}_2\text{O}]/[\text{H}_2]$) reaches 2.76×10^{-3} , corresponding to a twofold solar oxygen abundance. For deeper levels, the mixing ratio is maintained constant at this value⁹. Model 1 has one cloud

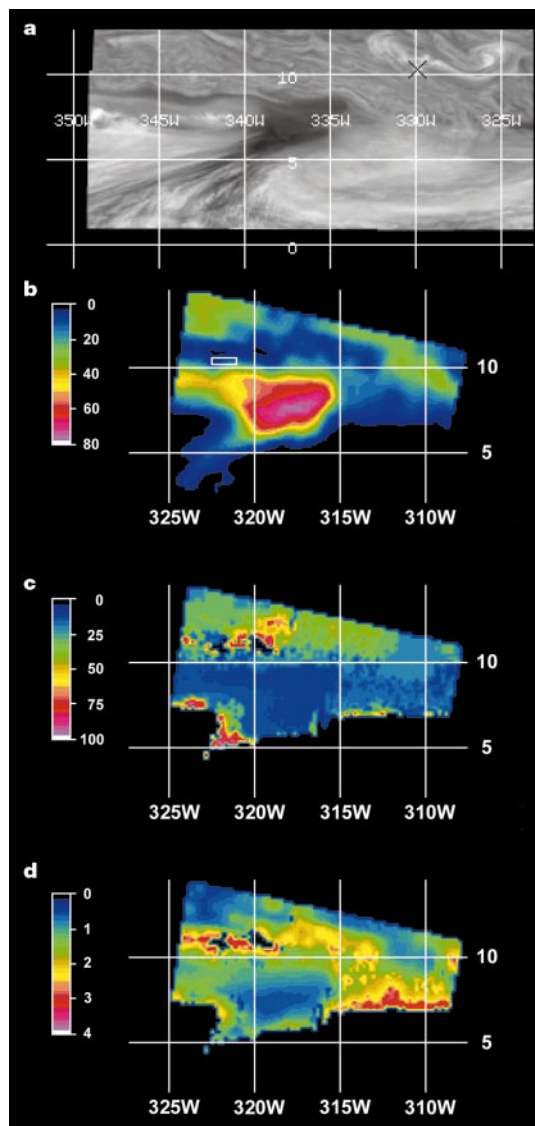


Figure 1 Comparison between the visible Solid State Imager (SSI) image, the 5- μm Near Infrared Mapping Spectrometer (NIMS) image, and results from the retrieval of water vapour and cloud opacity from the NIMS 5- μm spectra. **a**, SSI images acquired using the 0.756- μm continuum filter on 17 December 1996 at 17:13 UT (spatial resolution 30 km pixel⁻¹). A cloud system with high bright clouds at 329.9° W and 10.3° N can be identified (black cross). **b**, NIMS image at 5.052 μm taken on the night side of the planet on 20 December 1996 at 09:06 UT (spatial resolution 640 km pixel⁻¹). Units on the colour bar give radiance in $\mu\text{W cm}^{-2} \mu\text{m}^{-1}$. The shape of the hot spot (dark in the visible and bright at 5 μm) did not change significantly in the 64 h between the reflected sunlight and thermal observations. The positions of features visible in **a** can be calculated using the zonal velocities derived from a set of SSI images spanning 11 h (ref. 19). The bright cloud has an eastward (to the right) velocity of $\sim 40 \text{ ms}^{-1}$. When the NIMS thermal data were acquired, the centre of the bright cloud system was expected to be between 322.5° and 320.9° W (System III) and 10.4° N (rectangle). **c**, Map of the water vapour relative humidity as derived from the E4 NIMS 5- μm spectra. Units on the colour bar give the relative humidity in per cent. High relative humidities are required to fit spectra from the region near the predicted location of the bright cloud (Fig. 2). **d**, Map of the 1.55-bar cloud opacities as derived from the NIMS 5- μm spectra. Units on the colour bar give the optical depth of the cloud at 1.55 bar. High cloud opacities are derived in the humid area, consistent with the SSI observations of a thick cloud deck.

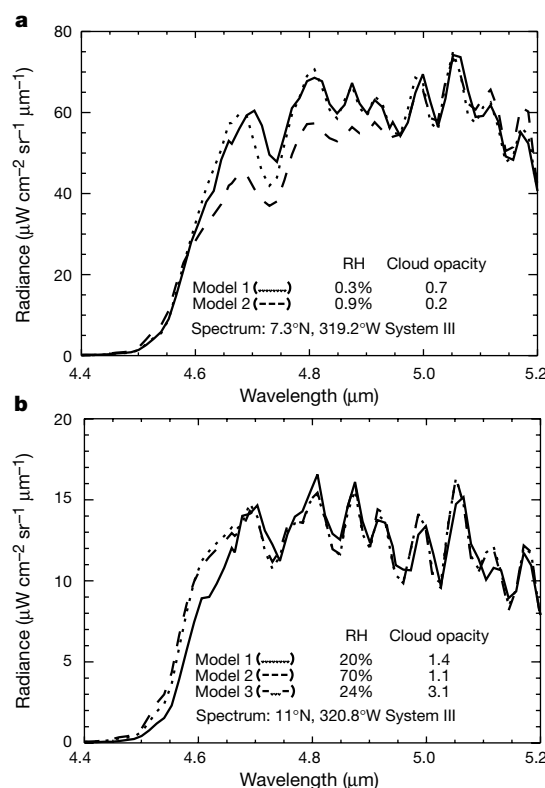


Figure 2 NIMS 5- μm spectra from a dry and a humid area. The 5- μm window of the jovian spectrum is free from strong absorptions of methane and ammonia. Radiation comes from 3 bar at 5.1 μm down to 8 bar at 4.65 μm . The overall continuum level is also sensitive to the total amount of cloud opacity above 2 bar. At NIMS spectral resolution, water dominates the absorption for wavelengths longer than 4.95 μm . At smaller wavelengths, phosphine dominates^{9,10}. **a**, Measured dry spectrum in the E4 hot spot area. The observed spectrum (solid line) was acquired at 319.2° W (System III) and 7.3° N and the 5.052- μm water band depth is 0.24. The values of the derived relative humidity and cloud optical depth (clouds above the 2-bar level) are listed for model 1 and 2. Only model 1 gives a good fit to the spectrum. Model 2, including an opaque cloud at 5.6 bar, retrieves higher water abundances, as explained in the text. **b**, Measured humid spectrum close to the E4 bright cloud. The observed spectrum (solid line) was acquired at 320.8° W (System III) and 11° N and the 5.052- μm water band depth is 0.34. All models (described in the text) provide good fits to the spectrum. The fact that model 3 gives similar results to model 1 for the RH confirms the validity of the non-scattering approximations applied in models 1 and 2. The difference in the estimated cloud opacities is due to the highly forward scattering particles of the 1.55-bar cloud in model 3, in contrast with the purely absorbing layer in model 1. The NIMS spectra have a noise level of about 5% on the radiance. This results in uncertainties of about 0.5 in the cloud opacity for opacities larger than 1, and a factor of about 3 for the relative humidity (models 1 and 2).

layer at 1.55 bar, similar to observations of the Galileo probe⁷, whereas model 2 includes an additional opacity source (the water cloud) at 5.6 bars, similar to predictions from thermodynamical models^{1,2} and some inferences from Voyager observations⁵. In addition, a correlated- k code including multiple scattering¹⁰ was used to confirm its validity, with one cloud layer at 1.55 bar (model 3). The atmospheric thermal structure is taken from Galileo probe *in situ* measurements¹¹. Cases were run assuming a range of RH and two models of vertical cloud structure.

Model 1 produces a good fit to hot spot spectra for an RH of less than $\sim 1\%$. Model 2, which includes the 5.6-bar cloud, produces a poor fit and rules out the existence of a deep cloud within the hot spot (Fig. 2). Outside the hot spot, both models produce similarly good fits. In all cases, however, model 2 requires a much greater RH than model 1. This non-uniqueness can be understood when studying the contribution functions. They are the kernel of the radiative transfer integral, that is, the variation of the gas opacity with altitude times the emission of each atmospheric layer. When opacity varies exponentially, as in the case of an atmosphere in hydrostatic equilibrium assumed here, the contribution functions peak at a certain level, which consequently contributes most of the outgoing radiation at a given wavelength. For small values of RH, the contribution functions peak near 6 bars, so that a cloud layer at 5.6 bars strongly affects the shape of spectra measured from orbit¹² (Fig. 2). When RH is more than $\sim 20\%$, the greatest contribution is from above the 5.6-bar level, so that the shape of the spectrum is not much affected by the presence of an opacity layer at this level. The effects of opacity at 5.6 bars are to 'cut off' contribution to the outgoing radiation from below this level, and thus require a higher water vapour concentration above this level to still fit the water band depth. Because it is not possible to unambiguously detect the presence of a water cloud in this region, the RH from model 1 should be taken as a lower limit.

Our primary finding is that extremely dry regions, such as the E4 hotspot with RH between 0.1% and 1%, exist in close proximity to humid areas, characterized by a RH of more than $\sim 20\%$. The presence of a deep water cloud is excluded in the dry regions. Although the precision of derived water vapour mixing ratios is limited by model assumptions, the above analysis shows that relatively dry and humid areas are easily distinguished using the observed 5.025- μm water band depth. Inside the hot spot the water band is shallow ($D = 0.2\text{--}0.28$), whereas outside the hot spot it is generally deeper ($D = 0.3\text{--}0.33$). At other locations it is very deep ($D = 0.33\text{--}0.4$). Clearly, local meteorology creates a more complex distribution of relative humidity than is expected solely from thermodynamical models^{1,2}.

Images at visible and near-infrared wavelengths from the SSI yield some insight into the local dynamics and cloud structure down to about 3 bars¹³. Using SSI images, deep (>3 bars) clouds have been found in the area surrounding the E4 bright cloud¹⁴. This bright cloud lies in a cyclonic shear zone, where the horizontal zonal flow is changing direction from about 100 ms^{-1} eastward at 6°N to about 25 ms^{-1} westward at 15°N (ref. 15). Cyclonic shear zones on Jupiter are characterized by chaotic cloud systems, outbursts of bright cloud material and, from recent Galileo results, the occurrence of lightning^{16,17}. When the bright cloud's position is corrected for wind advection, it coincides with a NIMS humid area (Fig. 1). Our spectroscopic results of that area greatly strengthen the case that the observed deep clouds are indeed convecting water clouds which result in lightning and the formation of bright cloud materials at higher altitudes. These widespread and energetic convection events may be a dominant source of energy for driving Jupiter's winds¹⁸. □

Received 8 October 1999; accepted 29 February 2000.

- Weidenschilling, S. J. & Lewis, J. S. Atmospheric and cloud structure of the jovian planets. *Icarus* **20**, 465–476 (1973).
- Atreya, S. K. *et al.* A comparison of the atmospheres of Jupiter and Saturn: deep atmospheric composition, cloud structure, vertical mixing, and origin. *Planet. Space Sci.* **47**, 1243–1262 (1999).

- Drossart, P. & Encrenaz, T. The abundance of water on Jupiter from the Voyager IRIS data at 5 microns. *Icarus* **52**, 483–491 (1982).
- Bjoraker, G., Larson, H. P. & Kunde, V. The abundance and distribution of water vapor in Jupiter's atmosphere. *Astrophys. J.* **311**, 1058–1072 (1986).
- Carlson, B. E., Lacy, A. A. & Rossow, W. B. The abundance and distribution of water vapor in the jovian troposphere as inferred from Voyager IRIS observations. *Astrophys. J.* **388**, 648–668 (1992).
- Niemann, H. B. *et al.* The composition of the jovian atmosphere as determined by the Galileo probe mass spectrometer. *J. Geophys. Res.* **103**, 22831–22845 (1998).
- Regent, B. *et al.* The clouds on Jupiter: results on the Galileo Jupiter mission probe nephelometer experiment. *J. Geophys. Res.* **103**, 22891–22910 (1998).
- Larson, H. P., Fink, U., Treffers, R. & Gautier, T. N. Detection of water vapour on Jupiter. *Astrophys. J.* **197**, L137–L140 (1975).
- Roos-Serote, M. *et al.* Analysis of Jupiter North Equatorial Belt hot spots in the 4–4 μm range from Galileo/near-infrared mapping spectrometer observations: measurements of cloud opacity, water and ammonia. *J. Geophys. Res.* **103**, 23023–23041 (1998).
- Irwin, P. *et al.* Cloud structure and atmospheric composition of Jupiter retrieved from Galileo near-infrared mapping spectrometer real-time spectra. *J. Geophys. Res.* **103**, 23001–23022 (1998).
- Seiff, A. *et al.* Thermal structure of Jupiter's atmosphere near the edge of a 5- μm hotspot in the north equatorial belt. *J. Geophys. Res.* **103**, 22857–22890 (1998).
- Roos-Serote, M., Drossart, P., Encrenaz, T., Carlson, R. W. & Leader, F. Constraints on the tropospheric cloud structure of Jupiter from spectroscopy in the 5- μm region: a comparison between voyager/IRIS, Galileo/NIMS, and ISO/SWS spectra. *Icarus* **137**, 315–340 (1999).
- Banfield, D. *et al.* Jupiter's cloud structure from Galileo imaging data. *Icarus* **135**, 230–250 (1998).
- Gierasch, P. *et al.* Cloud structure near convective locations on Jupiter. *Bull. Am. Astron. Soc.* **31**, 115 (1999).
- Limaye, S. S. Jupiter: New estimates of the mean zonal flow at the cloud level. *Icarus* **65**, 335–352 (1986).
- Little, B. *et al.* Galileo images of lightning on Jupiter. *Icarus* **142**, 306–323 (1999).
- Gierasch, P. *et al.* Observation of moist convection in Jupiter's atmosphere. *Nature* **403**, 628–630 (2000).
- Ingersoll, A. P., Gierasch, P. J., Banfield, D. & Vasavada, A. R. Moist convection as an energy source for the large-scale motions in Jupiter's atmosphere. *Nature* **403**, 630–632 (2000).
- Vasavada, A. R. *et al.* Jupiter's atmosphere: The Great Red Spot, equatorial region, and white ovals. *Icarus* **135**, 265–275 (1998).

Acknowledgements

This work was partly carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA, through the Jupiter System Data Analysis Program and the Galileo Project. M.R.-S. acknowledges financial support from PRAXIS XXI/FCT (Fundação para a Ciência e a Tecnologia), and the French Embassy/ICCTI (Instituto de Cooperação Científica e Tecnológica Internacional), Portugal. We thank P. Gierasch, D. Banfield and T. Entenaz for very constructive discussions, and J. Yoshimizu for help in making Fig. 1.

Correspondence and requests for materials should be addressed to M. R.-S. (e-mail: roos@oal.ul.pt).

Closing the spin gap in the Kondo insulator $\text{Ce}_3\text{Bi}_4\text{Pt}_3$ at high magnetic fields

Marcelo Jaime*, Roman Movshovich*, Gregory R. Stewart†, Ward P. Beyersmann‡, Mariano Gomez Berisso§, Michael F. Hundley*, Paul C. Canfield|| & John L. Sarrao*

*Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

†University of Florida, Gainesville, Florida 32611, USA

‡University of California, Riverside, California 92521, USA

§Centro Atómico Bariloche, Bariloche 8400, Argentina

||Ames Laboratory and Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, USA

Kondo insulator materials¹—such as CeRhAs , CeRhSb , YbB_{12} , $\text{Ce}_3\text{Bi}_4\text{Pt}_3$ and SmB_6 —are $3d$, $4f$ and $5f$ intermetallic compounds that have attracted considerable interest in recent years^{2–5}. At high temperatures, they behave like metals. But as temperature is reduced, an energy gap opens in the conduction band at the Fermi energy and the materials become insulating. This contrasts with other f -electron compounds, which are metallic at all

temperatures. The formation of the gap in Kondo insulators has been proposed to be a consequence of hybridization between the conduction band and the *f*-electron levels^{6,7}, giving a 'spin' gap. If this is indeed the case, metallic behaviour should be recovered when the gap is closed by changing external parameters, such as magnetic field or pressure. Some experimental evidence suggests that the gap can be closed in SmB₆ (refs 5, 8) and YbB₁₂ (ref. 9). Here we present specific-heat measurements of Ce₃Bi₄Pt₃ in d.c. and pulsed magnetic fields up to 60 tesla. Numerical results and the analysis of our data using the Coqblin–Schrieffer model demonstrate unambiguously a field-induced insulator-to-metal transition.

Previous experimental studies have indicated that the spin gap (or 'Kondo gap', with a value of Δ) can be closed in SmB₆ with applied pressure⁸ and magnetic field⁵, and that the gap in YbB₁₂ can be closed with an applied magnetic field⁹. (But recent data obtained on CeNiSn, described in the past as a Kondo insulator whose small spin gap can be closed by a moderate magnetic field, have shown this material to have a metallic, rather than insulating, ground state (see ref. 10 and references therein). Attempts have also been made in the past to observe the closure of the gap in Ce₃Bi₄Pt₃; these attempts involved measurements of electrical resistivity^{11,12} and magnetization¹³, in both d.c. magnets and in pulsed magnets driven by capacitor banks. While a large negative magnetoresistance was observed in fields up to 60 T at low temperatures and interpreted as an indication of closing of the gap¹², the magnetization is linear in magnetic field in the same field and temperature ranges.

In the light of these conflicting results, it has been suggested¹³ that the recovery of charge carriers observed in the transport properties in magnetic fields might be due to pockets of the Fermi surface where the gap is smaller. Another possibility is that the *g*-factor in Ce₃Bi₄Pt₃, which determines the energy scale of magnetic interactions, is smaller than expected (a larger magnetic field is needed to reach the range $g\mu_B H \approx \Delta$) and consequently the observed low-temperature properties are due simply to extrinsic charge carriers¹⁴. (Here *H* is the magnetic field, μ_B is the Bohr magneton and *J* is the electron's total angular momentum). Measurements of the specific heat in magnetic fields directly probe the evolution of the excitation spectrum, and of the Kondo gap in particular, and can therefore provide the key to understanding the physical origins of the very

distinctive ground-state properties of Kondo insulators, helping in the construction of a complete theoretical model.

The 60-tesla long-pulse (60TLP) magnet, at the National High Magnetic Field Laboratory, Los Alamos Pulsed Field Laboratory, is driven by a 1.4-GW synchronous power generator; this magnet produces a flat-top field for a period of greater than 100 ms at 60 T, and for longer time at lower fields. We have built a calorimeter out of plastic materials that enables us to perform heat-capacity measurements at temperatures between 1.4 and 20 K in this magnet¹⁵. During the magnetic field pulse, which lasts for about 2 s, the calorimeter can be regarded as thermally isolated, that is, in an adiabatic condition. We used a heat-pulse method, where a known amount of heat is delivered to the calorimeter using a chip resistor, to measure the heat capacity of flux-grown single-crystal samples of Ce₃Bi₄Pt₃. The data collected during a typical experiment on Ce₃Bi₄Pt₃ is displayed in Fig. 1, which shows that the calorimeter comes to thermal equilibrium both before and after the heat pulse is delivered within the magnetic field plateau. The temperature is measured during the entire field pulse with a Cernox bare chip resistance thermometer (Lakeshore Inc.), previously calibrated in d.c. field up to 30 T and in pulsed fields up to 60 T. The heat capacity of the sample is determined as the ratio of the heat delivered to the sample to the change in its temperature due to the heat pulse.

The results of the direct measurements of the specific heat $C(T)$ of a 44.85-mg single-crystal sample of Ce₃Bi₄Pt₃ (sample no. 2), performed in magnetic fields up to 60 T, are shown in Fig. 2. $C(T)/T$ is linear in T^2 , and the zero-temperature extrapolation γ_H increases from about 18 mJ mol⁻¹ K⁻² in zero field to close to 60 mJ mol⁻¹ K⁻² in a field of 60 T. All molar heat capacities reported here are per formula unit. The zero-field Sommerfeld coefficient is slightly larger than the values reported earlier⁶ (9–15 mJ mol⁻¹ K⁻²). This is consistent with measurements of the resistivity ratio in our samples $\rho(4\text{ K})/\rho(300\text{ K}) \approx 70$, which is somewhat lower than that previously reported¹². Both measurements indicate that the present sample contains a slightly larger concentration of defects or impurities. The observed increase in the heat capacity in fields equal to or larger than 30 T is in very good agreement with the reversible temperature change (due to the magnetocaloric effect¹⁵) observed during the ramp portions of the magnetic field pulse. We note the temperature drop during the ramp-up portion of the field profile in Fig. 1. In order to verify the calibration of our thermometry

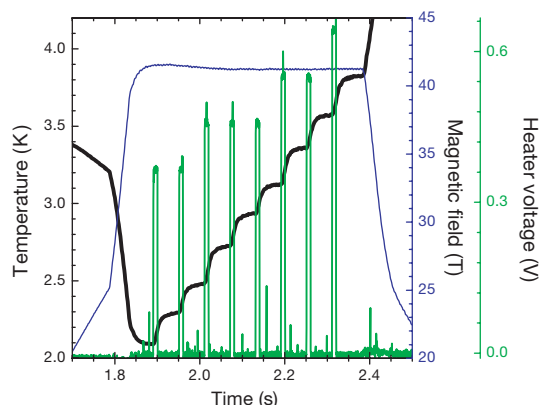


Figure 1 Heat-capacity experiment in the 60TLP magnet. The magnetic field profile is displayed in blue, the temperature of the Ce₃Bi₄Pt₃ (sample no. 1) in black, and eight (10-ms each) heat pulses delivered during the field plateau in green. The temperature decreases during the field ramp-up due to the adiabatic magnetization of the sample, reaches equilibrium before and after each heat pulse is delivered, and increases during the field ramp-down due to the adiabatic demagnetization of the sample.

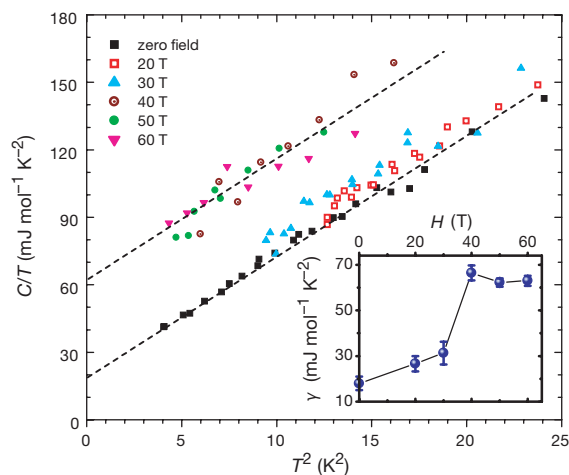


Figure 2 Specific heat in pulsed fields. Shown is C/T versus T^2 for Ce₃Bi₄Pt₃ single-crystal sample no. 2 in magnetic fields between zero and 60 T after subtraction of heat capacity of empty stage. Dashed lines are guides to the eye. Inset shows Sommerfeld coefficient $\gamma_H = C(H)/T_{T \rightarrow 0}$; obtained with a single-parameter fit (field-independent lattice contribution).

and correct operation of the calorimeter, we have also measured the specific heat of a 303-mg Si single crystal, where we do not expect any variation of specific heat with magnetic field. The magnetic field effect observed for this sample is no larger than $0.1\% \text{ T}^{-1}$, in accord with expectation.

In order to tell whether our results indicate the recovery of the metallic Kondo state in high fields, the increase in γ_H observed in $\text{Ce}_3\text{Bi}_4\text{Pt}_3$ needs to be put in perspective. Magnetic susceptibility¹⁴ and high-temperature neutron quasielastic line width^{7,16,17} measurements can be used to estimate a zero-field Kondo temperature of $T_K^0 = 240\text{--}320 \text{ K}$. In turn, the Sommerfeld coefficient for a metal with such T_K can be estimated using the expression for a single-impurity Kondo system¹⁸: $\gamma_0 = 3 \times 1.29 \pi R / 6T_K^0 = 53\text{--}70 \text{ mJ mol}^{-1} \text{ K}^{-2}$ (the pre-factor of 3 in the first expression accounts for the number of Ce atoms per formula unit and $R = 8.31 \text{ J mol}^{-1} \text{ K}^{-1}$ is the gas constant). This provides an upper bound for the high field γ_H in $\text{Ce}_3\text{Bi}_4\text{Pt}_3$, as it is expected that an external field will suppress correlations and induce a reduction in γ_H . Taking into account the effect of the applied magnetic fields within a single-impurity model¹⁹, our estimate of the Sommerfeld coefficient at 60 T is $\gamma_{60} = 51\text{--}66 \text{ mJ mol}^{-1} \text{ K}^{-2}$. Hundley *et al.*⁶ have measured the compound $\text{La}_3\text{Bi}_4\text{Pt}_3$ in zero field, and obtained $\gamma_0 = 27 \text{ mJ mol}^{-1} \text{ K}^{-2}$. This value in $\text{La}_3\text{Bi}_4\text{Pt}_3$, an isostructural metal where electronic correlations are absent, should be our lower bound limit on γ_{60} in the high field metallized state of $\text{Ce}_3\text{Bi}_4\text{Pt}_3$.

The inset in Fig. 2 shows γ_H for $\text{Ce}_3\text{Bi}_4\text{Pt}_3$ sample no. 2, in magnetic fields up to 60 T. The values for γ_H were obtained from a

single-parameter fit of the form $C(T) = \gamma_H T + \beta_H T^3$ (where $\gamma_H T$ and $\beta_H T^3$ are the electronic and phononic contributions in a magnetic field H , respectively), with the coefficient β_H of the lattice term fixed to its zero-field value. We see a sharp rise in γ_H between 30 and 40 T. The result of the fit suggests a saturation at a value of $\gamma_H^{\text{sat}} = 62 \pm 3 \text{ mJ mol}^{-1} \text{ K}^{-2}$ above 40 T. The strong enhancement of γ_H from its zero-field value, and the quantitative agreement with the estimate based on T_K for a metallic ground state of $\text{Ce}_3\text{Bi}_4\text{Pt}_3$, prove that we indeed crossed the phase boundary between the Kondo insulator and the Kondo metal.

The bandgap in $\text{Ce}_3\text{Bi}_4\text{Pt}_3$ should be observable in the temperature dependence of the specific heat at temperatures comparable to the gap, as the conduction band is depopulated. For lower magnetic fields available in d.c. superconducting magnets, we expect the value of the gap to remain high, and therefore, the experimental temperature range must be increased. A thermal relaxation technique was chosen to measure the specific heat of sample no. 2 in a superconducting magnet at temperatures between 4 and 60 K (ref. 20). The results of this experiment, C/T versus T , in zero field and 18 T, are displayed in Fig. 3a. We have subtracted the zero-field specific heat (C_0) from the specific heat in 18 T (C_{18}) to remove the non-field-dependent components, such as the phonon contribution, and divided the difference by the temperature. In this way we obtain the differential Sommerfeld coefficient at 18 T $\Delta\gamma_{18}(T)$ and display it in the inset of Fig. 3a. $\Delta\gamma_{18}$ is rapidly suppressed below 30 K, and remains finite at lower temperatures. This suggests a change in the magnitude of the gap, and a concomitant loss of charge carriers as the temperature is reduced. The error due to magnetoresistance in our temperature sensor (1% in T and 0.5% in C_{18} below 20 K) has been taken into account. The scatter (0.3% of C_0, C_{18}) reflects the typical reproducibility error of the experiment, and results in a relatively large error bar for the residual value $\Delta\gamma_{18|T=0} = 13 \pm 7 \text{ mJ mol}^{-1} \text{ K}^{-2}$.

We calculated the specific heat of a system with the following properties: narrow valence and conduction bands of equal width BW ($= 600 \text{ K}$), separated by a gap Δ , and a narrower impurity band of width w ($= 100 \text{ K}$) centred at the chemical potential μ , which is suggested theoretically²¹ as an explanation for the observed finite γ_0 (see Fig. 3a). In this model it was assumed that each band can accommodate $2N$ electrons, where N is the number of Ce atoms in the sample. We performed the computation for two cases: a fully open gap, and a reduced gap. The band widths remain constant, and the central band weight increases, when the gap is reduced. The spectral weight of the central band was chosen to reproduce the experimental γ_0 when the gap is open, and it is able to accommodate 0.96% of the total number of states in one band. Similar models (without the states in the gap) were used previously to simulate the zero-field electronic specific heat^{22,23} of $\text{Yb}_x\text{Lu}_{1-x}\text{B}_{12}$ and SmB_6 , and the NMR spin–lattice relaxation rate²⁴ of $\text{Ce}_3\text{Bi}_4\text{Pt}_3$. The difference between calculated specific heat in each case (with $\Delta = 155 \text{ K}$ and $\Delta = 220 \text{ K}$, simulating 18 T and zero field, respectively) is the solid line in the inset of Fig. 3a. The quantitative agreement with the data is evident.

The difference ΔC can remain finite at low temperatures only if some increase of the density of states at the chemical potential is allowed when the gap is reduced. The number of states we use for 18 T simulation is 1.7% of the states in the valence band. For comparison, a calculation with identical parameters and fixed (field independent) density of states at μ is the dashed line in the inset of Fig. 3a. The small but steady increase in γ_H observed in the pulsed-field experiment below 30 T (Fig. 2) and the finite $\Delta\gamma_{18|T=0}$ (inset in Fig. 3a) indicate that the effect of the applied field is to reduce the gap (Fig. 3b) as well as to increase the density of states inside the gap. The value obtained in our fit for the zero-field gap is in excellent agreement with the spin gap $\Delta_s = 250 \text{ K}$ observed in optical experiments²⁵. The gap value that we use to fit the specific-heat results is larger than the gap observed in transport experiments.

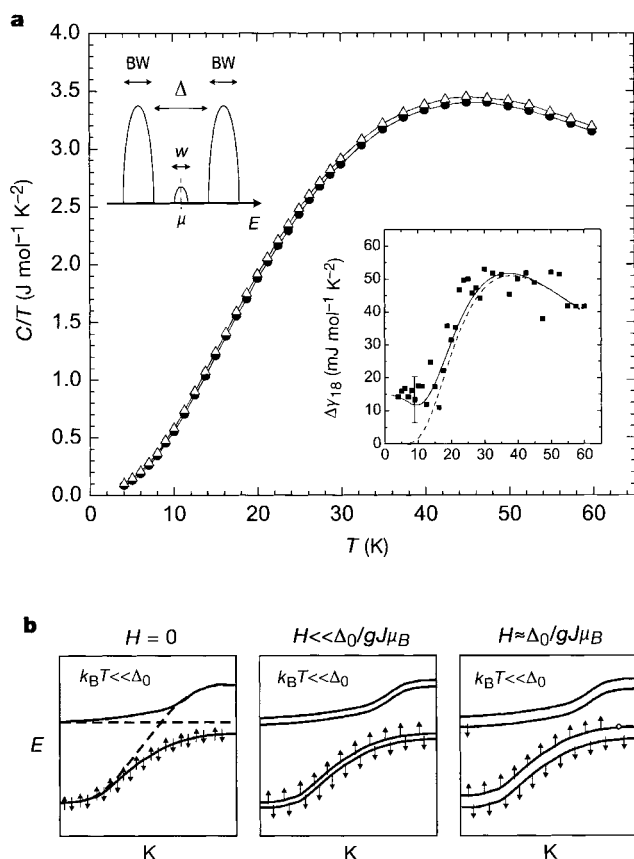


Figure 3 Specific heat measured in d.c. fields. **a**, C/T versus temperature T in zero field (solid circles) and 18 T (triangles), and diagram for the 'two narrow bands' model described in the text. Inset shows the differential Sommerfeld coefficient in a field of 18 T, $\Delta\gamma_{18} (= [C_{18} - C_0]/T)$, versus T . The lines are fits with the 'two narrow bands' model, with field-dependent (solid line) and field-independent (dashed line) impurity band at the chemical potential. **b**, Schematic band structure and magnetic field dependence in a Kondo insulator at low temperatures. Modified from ref. 1.

This discrepancy can be reconciled if one takes into account the impurity band centred at the chemical potential. Indeed, if the states in this band are localized or have small mobility at low temperatures, the conductivity of the material is very poor, as observed¹¹. As the temperature is increased, the relevant energy scale is the difference between the top of the impurity band and the bottom of the conduction band, which is 60 K in our fit and in very good agreement with the transport experiments^{11,12}. The reduction of the band gap observed in a field of 18 T is consistent with our low-temperature results that show the gap closing above 30 T. We also see that the magnitude of $\Delta\gamma_{18}$ at temperatures above 30 K, where the 18 T gap is expected to be effectively closed and the zero field gap to be open, is $50 \pm 5 \text{ mJ mol}^{-1} \text{ K}^{-2}$, in good agreement with the saturation value γ_H^{sat} for magnetic fields above 40 T at low temperatures.

These results of thermodynamic measurements add information to what we know about Kondo insulators from transport measurements in high fields^{5,9,12}, and should aid in the construction of a theoretical model for this class of strongly correlated compounds. Additionally, in the course of these studies, we have demonstrated the feasibility of direct specific-heat measurements in the extreme conditions of the pulsed magnetic fields produced by the 60TLP magnet at the National High Magnetic Field Laboratory. $\square$

Received 27 October 1999; accepted 14 March 2000.

1. Aeppli, G. & Fisk, Z. Kondo insulators. *Comments Cond. Mat. Phys.* **16**, 155–165 (1992).
2. Kumigashira, H. *et al.* Spectral evidence for pseudogap formation in Kondo insulators CeRhAs and CeRhSb. *Phys. Rev. Lett.* **82**, 1943–1946 (1999).
3. Susaki, T. *et al.* Temperature-dependent high-resolution photoemission study of the Kondo insulator YbB₁₂. *Phys. Rev. Lett.* **82**, 992–995 (1999).
4. Breuer, K. *et al.* Photoemission-study of the Kondo insulator Ce₃Bi₄Pt₃. *Europhys. Lett.* **41**, 565–570 (1998).
5. Cooley, J. C. *et al.* High field gap closure in the Kondo insulator SmB₆. *J. Supercond.* **12**, 171–173 (1999).
6. Hundley, M. F. *et al.* Hybridization gap in Ce₃Bi₄Pt₃. *Phys. Rev. B* **42**, 6842–6845 (1990).
7. Thompson, J. D. in *Transport and Thermal Properties of f-Electron Systems* (eds Fujii, H., Fujita, T. & Oomi, G.) 35–48 (Plenum, New York, 1993).
8. Moshchalkov, V. V. *et al.* SmB₆ at high-pressures: the transition from insulating to the metallic Kondo lattice. *J. Magn. Magn. Mat.* **47&48**, 289–291 (1985).
9. Sugiyama, K. *et al.* Field-induced metallic state in YbB₁₂ under high magnetic-field. *J. Phys. Soc. Jpn* **57**, 3946–3953 (1988).
10. Izawa, K. *et al.* Metallic ground state of CeNiSn. *Phys. Rev. B* **59**, 2599–2603 (1999).
11. Hundley, M. F. *et al.* Magnetoresistance of the Kondo insulator Ce₃Bi₄Pt₃. *Physica B* **186–188**, 425–427 (1993).
12. Boebinger, G. S., Passner, A., Canfield, P. C. & Fisk, Z. Studies of the Kondo insulator Ce₃Bi₄Pt₃ in 61-T pulsed magnetic-fields. *Physica B* **211**, 227–229 (1995).
13. Modler, R. in *Physical Phenomena at High Magnetic Fields-III* (eds Fisk, Z., Gor'kov, L. & Schrieffer, R.) 154–159 (World Scientific, Singapore, 1999).
14. Hundley, M. F. *et al.* Evidence for a coherence gap in Ce₃Bi₄Pt₃. *Physica B* **171**, 254–257 (1991).
15. Jaime, M. *et al.* in *Physical Phenomena at High Magnetic Fields-III* (eds Fisk, Z., Gor'kov, L. & Schrieffer, R.) 148–153 (World Scientific, Singapore, 1999).
16. Severing, A. *et al.* Gap in the magnetic excitation spectrum of Ce₃Bi₄Pt₃. *Phys. Rev. B* **44**, 6832–6837 (1991).
17. Hewson, A. C. *The Kondo Problem to Heavy Fermions* (Cambridge Univ. Press, Cambridge, 1993).
18. Rajan, V. T. Magnetic-susceptibility and specific-heat of the Coqblin-Schrieffer model. *Phys. Rev. Lett.* **51**, 308–311 (1983).
19. Schotte, K. D. & Schotte, U. Interpretation of Kondo experiments in a magnetic-field. *Phys. Lett. A* **55**, 38–40 (1975).
20. Bachmann, R. *et al.* Heat capacity measurements on small samples at low temperatures. *Rev. Sci. Instrum.* **43**, 205–214 (1972).
21. Schlottmann, P. Impurity bands in Kondo insulators. *Phys. Rev. B* **46**, 998–1004 (1992).
22. Iga, F., Kasaya, M. & Kasuya, T. Specific-heat measurements of YbB₁₂ and Yb₂Lu_{1-x}B_{12-x}. *J. Magn. Magn. Mat.* **76&77**, 156–158 (1988).
23. Mandrus, D. *et al.* Low-temperature thermal-expansion of SmB₆: evidence for a single-energy scale in the thermodynamics of Kondo insulators. *Phys. Rev. B* **49**, 16809–16812 (1994).
24. Reyes, A. P. *et al.* Bi-209 NMR and NQR investigation of the small-gap semiconductor Ce₃Bi₄Pt₃. *Phys. Rev. B* **49**, 16321–16330 (1994).
25. Degiorgi, L. The electrodynamic response of heavy-electron compounds. *Rev. Mod. Phys.* **71**, 687–734 (1999).

Acknowledgements

We thank Z. Fisk, J. D. Thompson and P. Schlottmann for discussions; J. Kim for his assistance with the thermometry calibration in the 30 T d.c. magnet at the NHMFL/Tallahassee; and D. Rickel, C. Mielke, J. Betts, J. Schillig, J. Sims and M. Pacheco for technical assistance and operation of the 60TLP magnet.

Correspondence and requests for materials should be addressed to M.J. (e-mail: mjaime@lanl.gov)

Relaxation in polymer electrolytes on the nanosecond timescale

Guomin Mao*, Ricardo Fernandez Perea*, W. Spencer Howells†, David L. Price* & Marie-Louise Saboungi*

* Argonne National Laboratory, Argonne, Illinois 60439, USA

† Rutherford-Appleton Laboratory, Chilton, Oxfordshire OX11 0QX, UK

The relation between mechanical and electrical relaxation in polymer/lithium-salt complexes is a fascinating and still unresolved problem in condensed-matter physics¹, yet has an important bearing on the viability of such materials for use as electrolytes in lithium batteries. At room temperature, these materials are biphasic: they consist of both fluid amorphous regions and salt-enriched crystalline regions. Ionic conduction is known to occur predominantly in the amorphous fluid regions. Although the conduction mechanisms are not yet fully understood², it is widely accepted that lithium ions, coordinated with groups of ether oxygen atoms on single or perhaps double polymer chains, move through re-coordination with other oxygen-bearing groups^{3,4}. The formation and disruption of these coordination bonds must be accompanied by strong relaxation of the local chain structure. Here we probe the relaxation on a nanosecond timescale using quasielastic neutron scattering, and we show that at least two processes are involved: a slow process with a translational character and one or two fast processes with a rotational character. Whereas the former reflects the slowing-down of the translational relaxation commonly observed in polyethylene oxide and other polymer melts, the latter appears to be unique to the polymer electrolytes and has not (to our knowledge) been observed before. A clear picture emerges of the lithium cations forming crosslinks between chain segments and thereby profoundly altering the dynamics of the polymer network.

The close relationship between ion transport and polymer relaxation has stimulated a considerable body of work on the dynamics of these systems. In particular, NMR line shapes and spin-lattice relaxation rates provide an effective probe of these dynamics on the microsecond to second timescale. For polymer electrolytes made of polyethylene oxide (PEO) with a number of

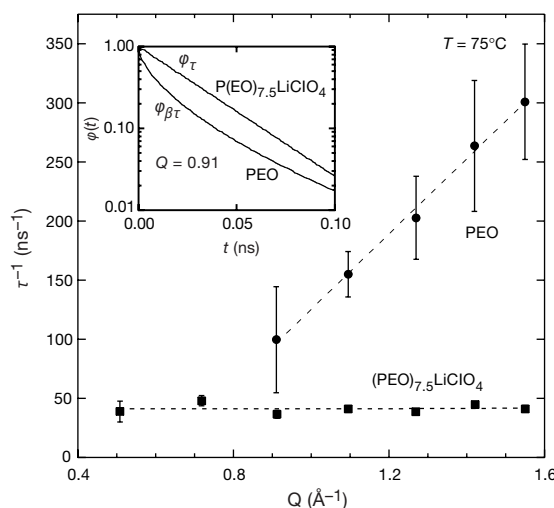


Figure 1 Variations of the inverse relaxation times τ^{-1} for pure PEO and P(EO)_{7.5}LiClO₄ electrolyte at 75 °C. Inset, behaviour of $\phi_{\beta}(t)$ and $\phi_{\tau}(t)$ for the two systems at $Q = 0.91 \text{ \AA}^{-1}$. Dashed lines are guides for the eye.

lithium salts, Li^+ was found to follow the segmental motions of the chain segments while the anions (ClO_4^- , BF_4^- , AsF_6^-) moved independently⁵. In contrast, the chains in pure PEO were found to be essentially stationary on the NMR timescale⁵.

The magnitude of the Li^+ diffusion constant in polymer electrolytes, $\sim 2 \times 10^{-7} \text{ cm}^2 \text{ s}^{-1}$ at 400 K (refs 3, 6), together with the distance between neighbouring ether oxygens of $\sim 3 \text{ \AA}$, implies a characteristic jump time of a few nanoseconds. In a computer simulation of $\text{P(EO)}_7\text{LiI}$ jump times for the cation were identified as 3–5 ns in the salt-rich compositions, while characteristic correlation times for the dihedral angle dynamics were somewhat shorter, $\sim 0.1 \text{ ns}$ (ref. 6). An experimental investigation of the dynamics of Li^+ diffusion—and the associated motions of polymer chain segments—appears, therefore, to require a probe with a dynamic range in the nanosecond timescale. Brillouin light scattering has been used to study the effects of both polymerization and salt complexing in $\text{P(PO)}_n\text{NaCF}_3\text{SO}_3$ but has a limited wavevector range⁷. (PPO is polypropylene oxide.)

Quasielastic neutron scattering (QENS) provides an especially powerful probe of polymer dynamics on the nanosecond timescale, with a significant dynamic range in both frequency and wavevector⁸. However, it has so far been used only on PPO-based systems which are complicated by the fast rotation of the methyl side groups on the timescale of interest^{9,10}. Here we report a study of pure PEO ($T_m \approx 60^\circ\text{C}$, where T_m is the melting point) and two electrolytes with an O(ether):Li ratio of 7.5:1. PEO-based systems are leading candidates for battery polymer electrolytes, and the perchlorate can be regarded as a prototypic salt that produces high ionic conductivity in the polymer–salt complex ($T_m \approx 50^\circ\text{C}$)¹¹; in contrast, PEO–LiTFSI (where TFSI refers to $\text{N(SO}_2\text{CF}_3)_2$) gives the best performance so far in terms of conductivity¹² and phase diagram (with T_m below room temperature)¹¹. The structures of both electrolytes have been measured using neutron diffraction with ^6Li – ^7Li isotope substitution (ref. 13; G.M. *et al.*, unpublished results).

We prepared sample films by stepwise drying solutions of the PEO (together with the salt for the electrolytes) in acetonitrile in a high-vacuum oven to 10^{-6} torr. Each film was loaded into a flat cell with thin aluminium windows sealed by Viton O-rings. QENS measurements were performed on the IRIS spectrometer at the pulsed spallation neutron source ISIS, Rutherford Laboratory, covering a wave vector Q range from 0.5 to $1.6 \text{ \AA}^{-1}$ with an energy resolution $\Delta E_{\text{res}} = 0.015 \text{ meV}$, and an energy transfer window

$\pm E_{\text{win}} = 0.4 \text{ meV}$. Data presented here were collected at 75°C in the liquid state to avoid complications arising from a biphasic system.

The scattering intensities $I(\omega)$ exhibited different behaviour for each of the samples examined; $\hbar\omega$ is the energy transfer. For PEO, $I(\omega)$ showed a broadening typical of that observed in molten polymers¹⁴ represented by the following expression:

$$I(\omega) = B\varphi_{\beta\tau}(\omega) + C \quad (1)$$

where B is a Q -dependent coefficient, and C accounts for very fast dynamical processes (for example, vibrations) and instrumental background. (The value of the coefficient C is of the same magnitude as the statistical errors, less than 1% of the total intensity.) The spectral function, $\varphi_{\beta\tau}(\omega)$, is the Fourier transform of the stretched exponential function, $\varphi_{\beta\tau}(t) = [1 - t/\tau]^\beta$, where τ is the relaxation time, commonly used to describe a variety of relaxation effects in polymers¹⁵. A fit of equation (1) to the data gave an average value of the stretching factor β of ~ 0.61 with a standard deviation of 0.05, over the Q range measured.

For the polymer electrolytes, the scattering showed a quite different behaviour from PEO, and depended on the nature of the anion. In the case of PEO– LiClO_4 , $I(\omega)$ could be fitted by an expression of the form:

$$I(\omega) = A\delta(\omega) + B\varphi_\tau(\omega) + C \quad (2)$$

where the first term reflects a slow relaxation process that appears static on the timescale of the spectrometer (0.09 ns), and the spectral function is now a lorentzian function, the Fourier transform of a pure exponential, $\varphi_\tau(t) = \exp[-(t/\tau)]$. The behaviour of $\varphi_{\beta\tau}(t)$ and $\varphi_\tau(t)$, shown in the inset to Fig. 1 for $Q = 0.91 \text{ \AA}^{-1}$, illustrates this difference. Information about the rate and character of the different relaxation processes is provided by the Q -dependence of the inverse relaxation times τ^{-1} (Fig. 1) and the intensity coefficients A and B (Fig. 2). It is clear that τ^{-1} rises rapidly with Q for PEO, while it is almost independent of Q for the polymer electrolyte. The coefficient B for pure PEO, and the sum $(A+B)$ for the electrolyte, decay slowly with Q , suggestive of a Debye–Waller effect. In the case of PEO, the rapidly increasing value of τ^{-1} indicates that the relaxation is related to the translational diffusion of the chain segments. For the polymer electrolyte, the relatively constant value of τ^{-1} , together with an intensity (B) that increases with Q , suggests a rotational character for the fast relaxation.

In the case of PEO–LiTFSI, the scattering intensity could be fitted by an equation similar to (2), but two fast processes (B_1 and B_2) were now required as well as the slow one (A). The inverse relaxation times, τ_1^{-1} and τ_2^{-1} , shown in Fig. 3, again vary rather slowly with Q . A decreases with Q as in the case of LiClO_4 , while both B_1 and B_2 increase steadily with Q (Fig. 4). Again, the relatively constant values

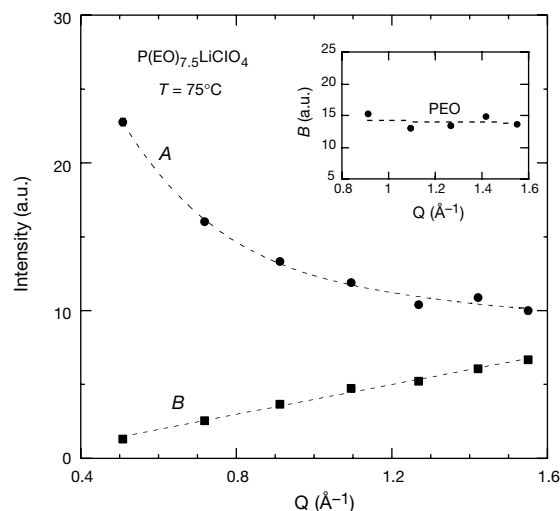


Figure 2 Coefficients of the slow (A) and fast (B) relaxation processes for the $\text{P(EO)}_{7.5}\text{LiClO}_4$ electrolyte at 75°C . A and B are defined in equation (2) in the text. Inset, B for pure PEO. Dashed lines are guides for the eye. a.u., arbitrary units.

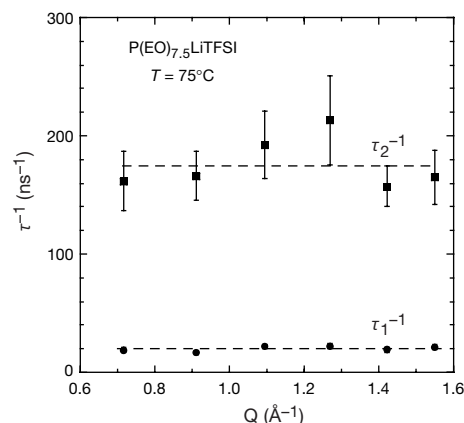


Figure 3 Inverse relaxation times τ_1^{-1} and τ_2^{-1} for the $\text{P(EO)}_{7.5}\text{LiTFSI}$ electrolyte at 75°C . Dashed lines are guides for the eye.

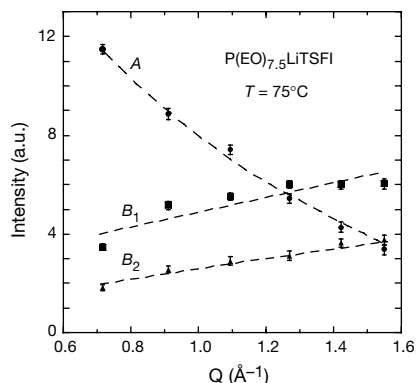


Figure 4 Coefficients of the slow (A) and fast (B_1 and B_2) relaxation processes for the P(EO)_{7.5}LiTFSI electrolyte at 75 °C. Dashed lines are guides for the eye.

of τ_1^{-1} and τ_2^{-1} , and intensities increasing with Q , suggest that both fast processes have a rotational character. It is tempting to ascribe the second process to the complexity of the TFSI anion: quantum-chemical calculations have identified two stable rotamers, C_1 and C_2 , in TFSI⁻, quite close in energy, with a very anharmonic potential for C_1/C_2 exchange¹⁶. It is possible that this exchange has a relaxational component that is being observed here, either directly or via coupling with the polymer host.

Because the scattering from lithium itself is an insignificant part of the total (0.05%), the different behaviour in PEO and the electrolytes reflects a profound change in the dynamics of the polymer chains induced by the addition of the salt. It is generally believed that, in relatively salt-rich polymer electrolytes, Li^+ ions form both inter-chain and intra-chain crosslinks, creating transient Li–O bonds which provide the mechanism for ion transport^{3,4}. Fast processes observed in the two complexes measured in the present experiment clearly arise from rapid conformational fluctuations of the chain segments between crosslinks; these fluctuations have a largely rotational character and correspond to the dihedral-angle relaxation observed in the computer simulations⁶. They must be associated with the making and breaking of the coordination bonds that enhances the Li^+ ion transport in the amorphous phases.

The slow processes that we observe in the polymer electrolytes reflect the slowing down of the translational motions of chain segments observed in PEO to a timescale that appears static on the spectrometer used. These studies are being extended to longer times (up to 2 ns) with neutron spin-echo spectrometry. Preliminary results confirm that the slow processes have a similar behaviour to the translational relaxation in pure PEO, in contrast to the rotational character of the fast motions (D. L. Price *et al.*, unpublished results). So it is clear that, on the nanosecond timescale, two or three different kinds of motion are involved in the lithium polymer electrolytes, and that the anion plays an important role in the dynamics. □

Received 10 October 1999; accepted 24 February 2000.

- Gray, F. M. *Polymer Electrolytes* (Materials Monographs, Royal Society of Chemistry, Cambridge, 1997).
- Berthier, C. *et al.* Microscopic investigation of ionic conductivity in alkali metal salts-poly(ethylene oxide) adducts. *Solid State Ionics* **11**, 91–95 (1983).
- Bruce, P. G. & Vincent, C. A. Polymer electrolytes. *J. Chem. Soc. Faraday Trans.* **89**, 3187–3203 (1993).
- Ratner, M. A. in *Polymer Electrolytes Review — 1* (eds McCallum, J. R. & Vincent, C. A.) 173–236 (Elsevier, London, 1987).
- Donoso, J. P. *et al.* Nuclear magnetic relaxation study of poly(ethylene oxide)-lithium salt based electrolytes. *J. Chem. Phys.* **98**, 10026–10036 (1993).
- Müller-Plathe, F. & van Gunsteren, W. F. Computer simulation of a polymer electrolyte: lithium iodide in amorphous poly(ethylene oxide). *J. Chem. Phys.* **103**, 4745–4756 (1995).
- Torrell, L. M., Jacobsson, P., Sidebottom, D. & Petersen, G. The importance of ion-polymer crosslinks in polymer electrolytes. *Solid State Ionics*, **53–56**, 1037–1043 (1992).
- Bee, M. *Quasielastic Neutron Scattering* (Adam Hilger, Bristol, UK, 1988).

- Carlsson, P. *et al.* Neutron-scattering studies of a polymer electrolyte, PPO–LiClO₄. *Solid State Ionics*, **113–115**, 139–147 (1998).
- Gabrys, B., Westerhuijs, P., Fishburn, J. & Barton, S. in *Quasielastic Neutron Scattering: Future Prospects on High-Resolution Inelastic Neutron Scattering* (eds Colmenero, J., Alegria, A. & Bermejo, F. J.) 221–226 (World Scientific, Singapore, 1994).
- Vallee, A., Besner, S. & Prud'homme, J. Comparative study of poly(ethylene oxide) electrolytes made with $\text{LiN}(\text{CF}_3\text{SO}_2)_2$, LiCF_3SO_3 and LiClO_4 : thermal properties and conductivity behaviour. *Electrochim. Acta* **37**, 1579–1583 (1992).
- Armand, M., Gorecki, W. & Andreani, R. Perfluorosulfonimide salts as solute for polymer electrolytes. *Proc. 2nd Int. Symp. on Polymer Electrolytes* (ed. Scrosati, B.) 91–97 (Elsevier, New York, 1990).
- Mao, G., Saboungi, M.-L., Price, D. L. & Armand, M. B. Structure of liquid PEO–LiTFSI electrolyte. *Phys. Rev. Lett.* (in the press).
- Arbe, A., Colmenero, J., Monkenbusch, M. & Richter, D. Dynamics of glass-forming polymers: “homogeneous” versus “heterogeneous” scenario. *Phys. Rev. Lett.* **81**, 590–593 (1998).
- Strobl, G. *The Physics of Polymers* (Springer, New York, 1996).
- Rey, L., Johansson, P., Lindgren, J., Lassegues, J. C., Grondin, J. & Servant, L. Spectroscopic and theoretical study of $(\text{CF}_3\text{SO}_2)_2\text{N}^-$ (TFSI⁻) and $(\text{CF}_3\text{SO}_2)_2\text{NH}$ (HTFSI). *J. Phys. Chem. A* **102**, 3249–3258 (1998).

Acknowledgements

This work was supported by The Divisions of Chemical Sciences and Materials Sciences, Office of Basic Energy Sciences, US Department of Energy. We thank M. B. Armand for pointing us to the PEO–LiTFSI electrolyte and for discussions, U. Buchenau for critical reading of the manuscript, the ISIS staff for experimental support, and D. S. Sivia for providing fitting routines.

Correspondence and requests for materials should be addressed to M.-L. S (e-mail: saboungi@anl.gov).

Non-fluorous polymers with very high solubility in supercritical CO₂ down to low pressures

Traian Sarbu, Thomas Styranec & Eric J. Beckman

Chemical Engineering Department, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA

Liquid and supercritical carbon dioxide have attracted much interest as environmentally benign solvents¹, but their practical use has been limited by the need for high CO₂ pressures to dissolve even small amounts of polar, amphiphilic, organometallic, or high-molecular-mass compounds^{2–4}. So-called ‘CO₂-philes’ efficiently transport insoluble or poorly soluble materials into CO₂ solvent, resulting in the development of a broad range of CO₂-based processes, including homogeneous and heterogeneous polymerization, extraction of proteins and metals, and homogeneous catalysis^{5–11}. But as the most effective CO₂-philes are expensive fluorocarbons, such as poly(perfluoroether), the commercialization of otherwise promising CO₂-based processes has met with only limited success. Here we show that copolymers can act as efficient, non-fluorous CO₂-philes if their constituent monomers are chosen to optimize the balance between the enthalpy and entropy of solute–copolymer and copolymer–copolymer interactions. Guided by heuristic rules regarding these interactions, we have used inexpensive propylene and CO₂ to synthesize a series of poly(ether-carbonate) copolymers that readily dissolve in CO₂ at low pressures. Even though non-fluorous polymers are generally assumed to be CO₂-phobic, we expect that our design principles can be used to create a wide range of non-fluorous CO₂-philes from low-cost raw materials, thus rendering a variety of CO₂-based processes economically favourable, particularly in cases where recycling of CO₂-philes is difficult.

Drawing from recent studies of CO₂ mixture thermodynamics^{12–15}, we proposed that a CO₂-philic material should be a

copolymer where one of the monomers (M_1) provides high flexibility and high free volume to enhance the entropy of mixing. We use low glass-transition temperature (T_g) and steric parameter¹⁶ (σ) as evidence for high flexibility and free volume. In addition, polymers of M_1 should exhibit a low cohesive energy density, to render solute-solute interactions as weak as possible¹², promoting dissolution in CO_2 . The second monomer (M_2) contains Lewis base groups that are known¹³⁻¹⁵ to interact specifically with CO_2 , yet which do not contain both hydrogen-bond donors and acceptors (the presence of both would tend to lower flexibility, and grossly inflate both T_g and the cohesive energy density). Ideally (but not necessarily), enthalpic interactions between M_1 and M_2 should be unfavourable, producing the copolymer effect that would also help to ease dissolution in CO_2 .

As the number of favourable specific interactions (CO_2 - M_2) increases, the enthalpy of mixing may become more favourable but the entropy of mixing will surely drop. Increasing the fraction of M_2 in the chain may also hinder segmental motion, further reducing the entropy of mixing. Finally, increasing the number of M_2 in the chain will raise the cohesive energy density of the copolymer, and thus a point may be reached where further incorporation of such groups is an enthalpic detriment, rather than an advantage. Thus, we surmised that the correct number of co-monomers (much like a "bound" co-solvent) per chain would be that which provides a suitable balance between favourable cross-interactions, unfavourable increases to solute cohesive energy density, and entropy loss through decreased chain flexibility. If the relative proportions of M_1 and M_2 are chosen correctly, the miscibility pressures of

optimal copolymers in CO_2 should be lower than those of either homopolymer.

The carbonyl group was chosen as our model Lewis base, as previous work¹⁵ has shown that it interacts specifically with CO_2 . Although such interactions have been shown to be weak, their effects on phase behaviour can be dramatic. Kazarian, for example, has speculated¹⁵ that these specific interactions explain the high degree of swelling of polyacrylates by CO_2 . We have observed previously¹⁷ that addition of a few acetate-containing side chains to a silicone can reduce miscibility pressures in CO_2 by 2,500 pounds per square inch (p.s.i.), while addition of alkyl side chains raises miscibility pressures. Finally, simply capping the CO_2 -phobic hydroxyl terminus of a poly(propylene glycol) polymer (Fig. 1) with an acetate group lowers miscibility pressures significantly, more so than capping with an alkyl group.

To create a model non-fluorous CO_2 -phile with the requisite M_1 (low T_g) and M_2 (Lewis base) repeat units, we incorporated carbonyls into a polyether (T_g s typically 190–220 K) via copolymerization of CO_2 with propylene oxide (PO). Previous work had shown that reaction of PO and CO_2 typically produces only cyclic propylene carbonate; so new sterically hindered aluminium catalysts were developed to synthesize the ether-carbonate copolymers (see Methods). The composition of the copolymers (Fig. 2) generated using these catalysts could be varied between a homopolymer of PO (100% ether) to a completely alternating copolymer of CO_2 and PO (100% polycarbonate).

Copolymers of PO and CO_2 proved to be very CO_2 -philic; Fig. 3a shows that a 250-repeat-unit PO/ CO_2 copolymer with 15.4%

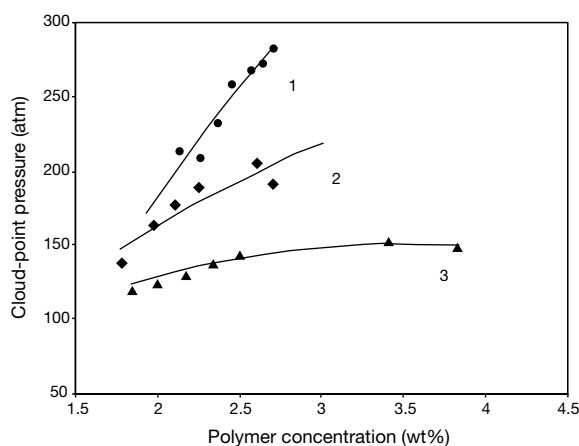


Figure 1 Addition of only one Lewis base group (carbonyl) to a polyether can significantly lower miscibility pressures in CO_2 . Shown is the phase behaviour of poly(propylene glycol) diol (curve 1), poly(propylene glycol) monobutyl ether (curve 2) and poly(propylene glycol) acetate (curve 3), each with 21 repeat units, in CO_2 at 22 °C

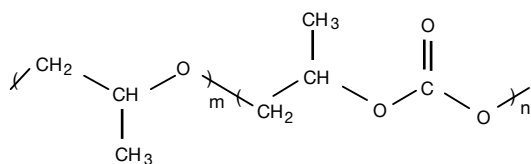


Figure 2 Structure of the CO_2 -philic poly(ether-carbonate) copolymers. Sterically hindered aluminium catalysts produce ether-carbonate copolymers from propylene oxide (PO) and CO_2 , where the composition can be varied from 100% polyether to 100% carbonate (completely alternating polymerization of CO_2 and oxirane).

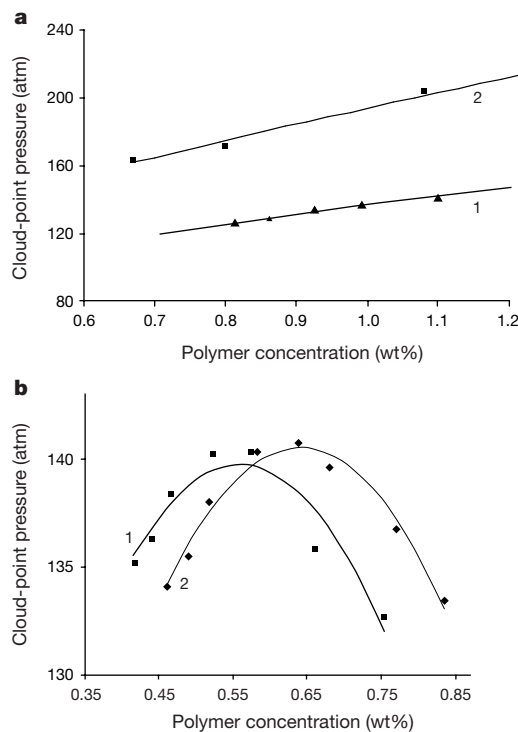


Figure 3 Miscibility of different CO_2 -philic polymers in CO_2 . **a**, An ether-carbonate copolymer will dissolve in CO_2 at lower pressures than a fluorinated polyether. Shown is a comparison of the phase behaviour (miscibility pressures in CO_2 at 22 °C) of a copolymer of PO and CO_2 (curve 1; 250 repeat units, 15.4 mol.% carbonate repeats) with a poly(perfluoropropylene oxide) polymer (curve 2; Krytox oil, DuPont, fractionated to give material with 175 repeat units). **b**, Copolymers of cyclohexene oxide (CHO) and CO_2 with relatively small amounts of carbonate will dissolve in CO_2 at pressures of 140 bar and below, although homopolymers of CHO of the same chain length are essentially insoluble. Shown is a comparison of the phase behaviour of CHO- CO_2 copolymers (curve 1, 124 repeat units, 8.8% carbonate; curve 2, 88 repeat units, 2.3% carbonate).

carbonate units exhibits lower miscibility pressures than a fluoro-ether—poly(hexafluoropropylene oxide)—whose chain length is significantly lower (175 repeats). We note that a PO homopolymer of 250 repeat units exhibits miscibility pressures beyond the capacity of our instrument. Clearly, addition of the Lewis base carbonyl groups contributes to this result, but, in addition, the literature shows that σ for polycarbonates is lower than that for aliphatic polyethers¹² (indeed, polycarbonates have some of the lowest σ s reported). Consequently, it appears that by incorporating the carbonyl group (via formation of the main chain carbonate linkage), we have both increased the degree of flexibility of the molecule (and thus the entropy of mixing) and enhanced the enthalpy through addition of the carbonyl groups. We have observed this effect for other oxirane/CO₂ copolymers as well—Fig. 3b shows that copolymers of cyclohexene oxide (CHO) and CO₂ exhibit miscibility pressures of 140 bar and less (homopolymers of CHO of such chain lengths exhibit miscibility pressures beyond the limits of our equipment). The data in Fig. 3b shows that only a small percentage of carbonate is required to achieve very marked improvements in phase behaviour, and that an increased percentage of carbonate compensates for an increase in chain length. Finally, a 103-repeat-unit copolymer of ethylene oxide (EO) and CO₂ (34% carbonate) exhibits the same miscibility pressures as a 16-repeat-unit homopolymer of EO, whereas an EO homopolymer of 103 repeat units is essentially insoluble in CO₂ (ref. 13).

We have proposed that the miscibility pressures of a copolymer exhibiting the optimal balance between the two monomers should be lower than either of the two homopolymers. Indeed, we have found that this is true for ether-carbonate copolymers generated from CO₂ and PO. Comparing the phase behaviour of a series of polymers, all with a chain length of 25 repeat units, we observed that a 100% carbonate polymer exhibits miscibility pressures beyond the limits of our instrument, while a 40% carbonate copolymer exhibits cloud-point pressures (see Methods) below that of a 100% ether (PO) polymer. We have observed that the optimal carbonate level in these ether-carbonate copolymers varies as the oxirane employed (EO, PO, epichlorohydrin, cyclohexene oxide) varies.

We suggest that these polymers could be used in CO₂-based applications where fluorinated materials are predominantly employed, including use as building blocks for surfactants. The EO/CO₂ copolymer mentioned above (103 repeats, 34% carbonate), for example, readily dissolves in water and CO₂—unusual behaviour for a polymer. When added to a biphasic mixture of water and CO₂ (1.2 vol.%, or 2 mM polymer), an emulsion forms upon

stirring at 340 bar, but settles after only an hour. Although the EO/CO₂ copolymer is apparently not an adequate surfactant, this material can be used as a building block for a CO₂-soluble amphiphile. We have previously used hydroxy-terminated fluoroethers and silicones as building blocks for CO₂-philic chelating agents and surfactants^{18,19}, and quenching of CO₂/oxirane polymerizations (mono-functional catalysts) with methanolic HCl readily produces hydroxy-terminated polymers. When an EO/CO₂ copolymer (57 repeat units, 8% carbonate) is modified with maleic anhydride, the resulting carboxy-functional polymer (1 vol.%, or 8.5 mM polymer) forms an emulsion in a CO₂/water mixture that is stable overnight.

Because the polymerization of CO₂ and oxiranes by our sterically hindered aluminium catalysts proceeds via a sequential insertion at the unhindered Al–O bond, it is relatively easy to prepare blocky copolymers for use as CO₂-soluble amphiphiles. We generated a triblock polymer where the end blocks are CO₂/CHO copolymers (25 repeats) and the middle block is hydrophilic polyethylene oxide (7 EO repeats). This surfactant is soluble in CO₂ at moderate pressures (Fig. 4), and when added (1 vol.%, or 1.8 mM polymer) to a mixture of water and CO₂, forms an emulsion upon stirring that is stable for hours. We observed that the emulsion was more stable at pressures between 170 and 350 bar than at pressures above 400 bar. We speculate that we are observing the upper critical micelle density observed²⁰ for fluoroacrylate-containing block copolymers. In our situation, the EO block, while relatively CO₂-phobic, can be solubilized by CO₂ at higher pressures. □

Methods

Catalyst preparation

Catalysts were prepared in monofunctional (R₁–O)₂–Al–O–R₂ and difunctional (R₁–O)₂–Al–O–R₃–O–Al–(O–R₁)₂ forms, where R₁ was triphenyl methyl, 2,6 *t*-butyl, 4-methyl phenyl, or fenchyl; R₂ was typically isopropyl, and R₃ was either isopropyl or (CH₂CH₂O)_x. Note that polymerization occurs exclusively at the non-sterically hindered Al–O bond, and thus at one position for the monofunctional catalyst and two positions for the difunctional version. Thus, the triblock polymer described above was generated using a difunctional catalyst where R₃ is (EO)₇. All catalysts were synthesized under argon, in glass flasks that had been heated to 200 °C, evacuated, then flushed with inert gas three times. They were synthesized by the modification of alkyl aluminium, in two steps, with tri(phenyl)methanol or a sterically hindered phenol, then with an appropriate alcohol or glycol. The ²⁷Al-NMR spectra of the catalysts showed that the sterically hindered aluminium catalysts exhibit only pentacoordination. In a typical experiment 10 ml of triisobutyl aluminium, as a 1.0 M solution in toluene, is reacted with 5.207 g (0.02 mol) of tri(phenyl)methanol at 40 °C for 2–4 h, and then 0.77 ml (0.01 mol) isopropanol is added dropwise, and the mixture stirred for two hours. The suspension of catalyst in toluene is cooled to room temperature, the solvent is removed with a syringe and the catalyst washed with a small amount of dried toluene, and after the procedure is repeated, the remaining solvent is removed in vacuum at 50 °C.

Polymerization

Polymerization of PO and CO₂ was performed in a 35 ml high-pressure reactor (manufactured at the University of Pittsburgh) equipped with magnetic stirrer and pressure and temperature indicators. Before the experiment the reactor is heated to 200 °C, evacuated, and then cooled to room temperature under an argon blanket. The desired amount of solid tri(phenyl)methoxy aluminium catalysts were introduced in the reactor under argon blanket, then the reactor was again evacuated for 15–20 min, then flushed with argon. The aluminium phenoxide-type catalysts were introduced with a syringe. PO was added alone using a syringe, or as a mixture with CO₂ using a high-pressure syringe pump (High Pressure Equipment Co.). After injection of the reagents, the reactor was isolated and heated to the prescribed temperature (40–60 °C). After the desired time the pressure was slowly released, and the reaction was then terminated with methanolic hydrochloric acid. The polymer solution was filtered using a membrane filter with a 0.45-μm pore size (Schleicher & Schuell), and the solvent was then removed under vacuum. Molecular masses and the molecular mass distribution of the polymers were measured using a Waters 150 CV gel permeation chromatograph, equipped with 10⁵, 10⁴, 500 and 100 Å ultrastaygel columns, where tetrahydrofuran was used as eluent and calibration was performed using polystyrene standards. The amount of carbonate incorporated into the polymer was determined from the integrals of the ether protons peaks at approximately 3.4–3.6 p.p.m. and the carbonate bound protons at 4.8 p.p.m. from ¹H-NMR.

Polymerizations using these aluminium catalysts are 'living' in character, because the molecular mass distributions are typically 1.1 and the absolute molecular mass is governed by the ratio of monomer to initiator. The extent to which CO₂ is incorporated into the polymer is a function of temperature, pressure and catalyst type. Reactions between PO

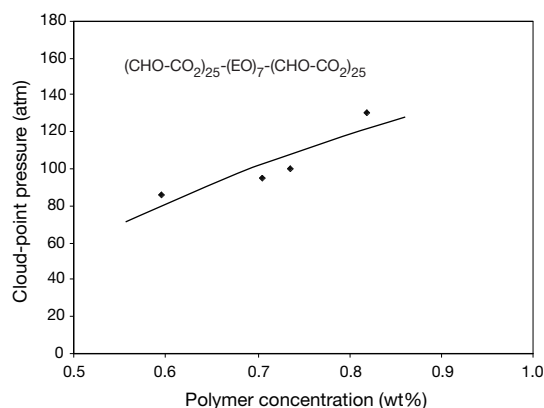


Figure 4 Triblock surfactants generated from polyethylene glycol, cyclohexene oxide (CHO), and carbon dioxide are soluble in CO₂ at low pressures. Shown is the phase behaviour of triblock terpolymer where the end blocks are generated from CHO and CO₂, while the middle block is a seven-repeat-unit ethylene oxide (EO) oligomer.

and CO₂ (at 40–60 °C) go to completion in less than 24 h. Alcohols can be used as chain transfer agents, reducing molecular mass without reduction in polymerization rate.

Measuring phase behaviour

The phase behaviour of the model polymers in CO₂ was measured using a high-pressure, variable-volume view cell (D.B. Robinson & Assoc.). This cell is a quartz tube containing a floating piston; the entire tube assembly is encased in a windowed, stainless-steel vessel capable of supporting pressures up to 500 bar. A known amount of solute is added to the view-cell chamber above the piston; the chamber is then sealed and charged with liquid CO₂. The pressure on the sample is altered at constant composition via injection of silicone oil, which moves the floating piston higher in the quartz tube. Initially, the pressure is raised to the point where a clear, single-phase solution is obtained. The phase boundary (cloud point, or miscibility pressure) is found subsequently by slowly lowering the pressure until the solution becomes cloudy^{21,22}.

Emulsion experiments

The ability of copolymers to form emulsions in CO₂/water mixtures was observed using a high-pressure, windowed reactor previously employed for conducting dispersion polymerizations in CO₂ (ref. 23). Water (28% by volume) and surfactant were charged to the reactor, which was then sealed and charged with CO₂ to the desired pressure. Stirring was accomplished using a high intensity, 3-blade mixer (600 r.p.m.).

Received 10 November 1999; accepted 6 March 2000.

- Eckert, C. A., Knutson, B. L. & DeBenedetti, P. G. Supercritical fluids as solvents for chemical and materials processing. *Nature* **373**, 313–318 (1996).
- Consani, K. A. & Smith, R. D. Observation on the solubility of surfactants and related molecules in carbon dioxide at 50 °C. *J. Supercrit. Fluids* **3**, 51–65 (1990).
- O'Shea, K. E., Kirmse, K. M., Fox, M. A. & Johnston, K. P. Polar and hydrogen-bonding interactions in supercritical fluids. Effects on the tautomeric equilibrium of 4-(phenylazo)-1-naphthol. *J. Phys. Chem.* **95**, 7863–7867 (1991).
- Johnston, K. P. & Lemert, R. M. In *Encyclopedia of Chemical Processing and Design* Vol. 56 (ed. McKetta, J. J.) 1–45 (Dekker, New York, 1996).
- Harrison, K., Goveas, J., Johnston, K. P. & O'Rear, E. A. Water-in-carbon dioxide microemulsions with a fluorocarbon-hydrocarbon hybrid surfactant. *Langmuir* **10**, 3536–3541 (1994).
- DeSimone, J. M., Guan, Z. & Elsbend, C. S. Synthesis of fluoropolymers in supercritical carbon dioxide. *Science* **267**, 945–947 (1992).
- Hsiao, Y. L., Maury, E. E., DeSimone, J. M., Mawson, S. M. & Johnston, K. P. Dispersion polymerization of methyl-methacrylate stabilized with poly(1,1-dihydroperfluorooctyl acrylate) in supercritical carbon dioxide. *Macromolecules* **28**, 8159–8166 (1995).
- Ghenciu, E. G., Russell, A. J., Beckman, E. J., Steele, L. & Becker, N. T. Solubilization of subtilisin in CO₂ using fluoroether-functional amphiphiles. *Biotech. Bioeng.* **58**, 572–580 (1998).
- Johnston, K. P. *et al.* Water in carbon dioxide microemulsions: An environment for hydrophiles including proteins. *Science* **271**, 624–626 (1996).
- Yazdi, A. V. & Beckman, E. J. Design of highly CO₂-soluble chelating agents. 2. Effect of chelate structure and process parameters on extraction efficiency. *Ind. Eng. Chem. Res.* **36**, 2368–2374 (1997).
- Jessop, P. G., Ikariya, T. & Noyori, R. Homogeneous catalysis in supercritical fluids. *Chem. Rev.* **99**, 475–494 (1999).
- O'Neill, M. L. *et al.* Solubility of homopolymers and copolymers in carbon dioxide. *Ind. Eng. Chem. Res.* **37**, 3067–3079 (1998).
- Rindfleisch, F., DiNoia, T. P. & McHugh, M. A. Solubility of polymers in supercritical CO₂. *J. Phys. Chem.* **100**, 15581–15587 (1996).
- Meredith, J. C., Johnston, K. P., Seminario, J. M., Kazarian, S. G. & Eckert, C. A. Quantitative equilibrium constants between CO₂ and Lewis bases from FTIR spectroscopy. *J. Phys. Chem.* **100**, 10837–10848 (1996).
- Kazarian, S. G., Vincent, M. F., Bright, F. V., Liotta, C. L. & Eckert, C. A. Specific intermolecular interaction of carbon dioxide with polymers. *J. Am. Chem. Soc.* **118**, 1729–1736 (1996).
- Brandrup, J., Immergut, E. H. & Grulke, E. A. (eds) *Polymer Handbook* 4th edn Ch. VII, 47–68 (Wiley, New York, 1999).
- Fink, R., Hancu, D., Valentine, R. & Beckman, E. J. Toward the development of "CO₂-philic" hydrocarbons. 1. Use of side-chain functionalization to lower the miscibility pressure of polydimethylsiloxanes in CO₂. *J. Phys. Chem. B* **103**, 6441–6444 (1999).
- Hoefling, T. A., Enick, R. M. & Beckman, E. J. Microemulsions in near critical and supercritical CO₂. *J. Phys. Chem.* **95**, 7127–7129 (1991).
- Li, J. & Beckman, E. J. Affinity extraction into CO₂. 2. Extraction of heavy metals into CO₂ low-pH aqueous solutions. *Ind. Eng. Chem. Res.* **37**, 4768–4773 (1998).
- Triolo, F. *et al.* Critical micelle density for the self-assembly of block copolymer surfactants in supercritical carbon dioxide. *Langmuir* **16**, 416–421 (2000).
- Hoefling, T. A., Newman, D. A., Enick, R. M. & Beckman, E. J. Effect of structure on the cloud point curves of silicone-based amphiphiles in supercritical carbon dioxide. *J. Supercrit. Fluids* **6**, 165–171 (1993).
- Newman, D. A., Hoefling, T. A., Beitle, R. R., Beckman, E. J. & Enick, R. M. Phase behavior of fluoroether-functional amphiphiles in supercritical carbon dioxide. *J. Supercrit. Fluids* **6**, 205–210 (1993).
- Lepilleur, C. & Beckman, E. J. Dispersion polymerization of methyl methacrylate in CO₂. *Macromolecules* **30**, 745–750 (1997).

Acknowledgements

We thank the US DOE, National Petroleum Technology Office, for their support of our CO₂ enhanced oil recovery research, and the US DOE, National Energy Technology Laboratory, for their support of our CO₂ well fracturing research.

Correspondence and requests for materials should be addressed to E.J.B. (e-mail: beckman@engr.pitt.edu).

Resolving the 'opal paradox' in the Southern Ocean

Philippe Pondaven*, Olivier Ragueneau*, Paul Tréguer*, Anne Hauvespre*, Laurent Dezileau† & Jean Louis Reyss†

* Université de Bretagne Occidentale UMR CNRS 6539, Institut Universitaire Européen de la Mer, Place Copernic, Technopôle Brest-Iroise, 29280 Plouzané, France

† Laboratoire des Sciences du Climat et de l'Environnement, Laboratoire mixte CNRS/CEA, 91191 Gif sur Yvette, France

In the Southern Ocean, high accumulation rates of opal—which forms by precipitation from silica-bearing solutions—have been found in the sediment in spite of low production rates of biogenic silica and carbon in the overlying surface waters. This so-called 'opal paradox' is generally attributed to a higher efficiency of opal preservation in the Southern Ocean than elsewhere^{1,2}. Here we report biogenic silica production rates, opal rain rates in the water column and opal sediment burial rates for the Indian Ocean sector of the Southern Ocean, which show that the assumed opal paradox is a result of underestimated opal production rates and overestimated opal accumulation rates. Our data thus demonstrate that the overall preservation efficiency of biogenic opal in this region is substantially lower than previously thought², and that it lies within a factor of two of the global mean³. The comparison of our revised opal preservation efficiencies for the Southern Ocean with existing values from the equatorial Pacific Ocean and the North Atlantic Ocean shows that spatial differences in preservation efficiencies are not the primary reason for the differences in sedimentary opal accumulation. The reconciliation of surface production rates and sedimentary accumulation rates may enable the use of biogenic opal in the reconstruction of palaeo-productivity when the factors that affect the Si/C ratio are better understood.

Comparison of hydrographic data from two ANTARES cruises (Southern Ocean—Joint Global Ocean Flux Study) performed in the Indian sector of the Southern Ocean along a 62° E transect, during early summer 1994 and late winter 1995, reveals marked seasonal nutrient depletion in some surface waters (Fig. 1): this depletion is typical of the Polar Front zone (PFZ), the permanently open ocean zone (POOZ), and the seasonal ice zone (SIZ). In addition, unaltered remnant winter water is observed at the depth of the temperature minimum (Fig. 1). The salinity in the remnant winter water layer is within 0.02–0.2% of that observed at the sea surface during winter. The same observation is valid for silicic acid and nitrate, with a concentration in remnant winter water being within 0.7–7% of that at the sea surface (Fig. 1). Thus we calculate the vertically integrated nutrient depletion during summer by reference to the remnant winter water concentrations^{4–6}, with an uncertainty ≤7%; we assume that the nutrient depletion by phytoplankton (referred to as ΔSi for silicic acid and ΔNO₃ for nitrate) occurs over a period of 90 days in this area, that is, from early November to early February⁵. We then derive an estimate of daily and annual production rates of Si, N and C from integrated nutrient depletion (see Methods).

Figure 2a illustrates the latitudinal variations of the summer/winter contrast in concentrations of dissolved silicic acid and nitrate in the Indian sector of the Southern Ocean. The nutrient depletion of surface waters reaches a latitudinal average of 13.31 ± 4.94 mmol Si m⁻³ and 3.01 ± 1.16 mmol NO₃ m⁻³ (±s.d.), illustrating the excess of silicic acid utilization compared to nitrate utilization in this region. Note that our estimate of nutrient depletion may be on the low side as further depletion may have occurred later in the season, after the ANTARES 2 cruise. The

excess of silicic acid utilization compared to nitrate utilization has been also documented for the Atlantic⁷ and Pacific⁸ sectors of the POOZ. The higher silicic acid depletion compared to nitrate depletion south of the PFZ may originate from the increase of the regenerated nitrogen production during summer^{5,9}, and from differential effects of light¹⁰ and nutrient limitation, especially Fe (ref. 11), on the Si/N uptake ratios.

Figure 2b indicates the vertically integrated nutrient depletion at the early summer ANTARES 2 stations. A significant latitudinal variation in the magnitude of the silicic acid depletion is seen. The depletion of silicic acid is highest in the POOZ, south of the PFZ, with an average of $2.31 \pm 0.49 \text{ mol Si m}^{-2}$ ($\pm$ s.d.) between 51°S and 58°S (Fig. 2b and Table 1). The latitudinal variations of the nitrate depletion are lower than for silicic acid, ranging from a minimum of

$0.14 \pm 0.07 \text{ mol NO}_3^- \text{ m}^{-2}$ within the SIZ to a maximum of $0.53 \pm 0.10 \text{ mol NO}_3^- \text{ m}^{-2}$ within the POOZ and the PFZ (Fig. 2b and Table 1). Table 1 gives an estimate of the average daily production rate of biogenic silica for the 90 days of the spring–summer period (see Methods). The production reaches a maximum of $30.3 \pm 6.0 \text{ mmol Si m}^{-2} \text{ d}^{-1}$ within the POOZ, and values which are 33% lower in the PFZ and 56% lower in the SIZ (Table 2). Conversely, the nitrogen production of $13.5 \pm 5.5 \text{ mmol N m}^{-2} \text{ d}^{-1}$ for the POOZ is 16% lower than in the PFZ, and 37% higher than in the SIZ (Table 2).

For both the PFZ and the SIZ, our estimates of the daily spring–summer Si and N production rates for 1994 are consistent with *in situ* measurements obtained for the Indian sector⁹ or in the Atlantic and Pacific sectors^{2,12}. For the POOZ however, the estimated daily

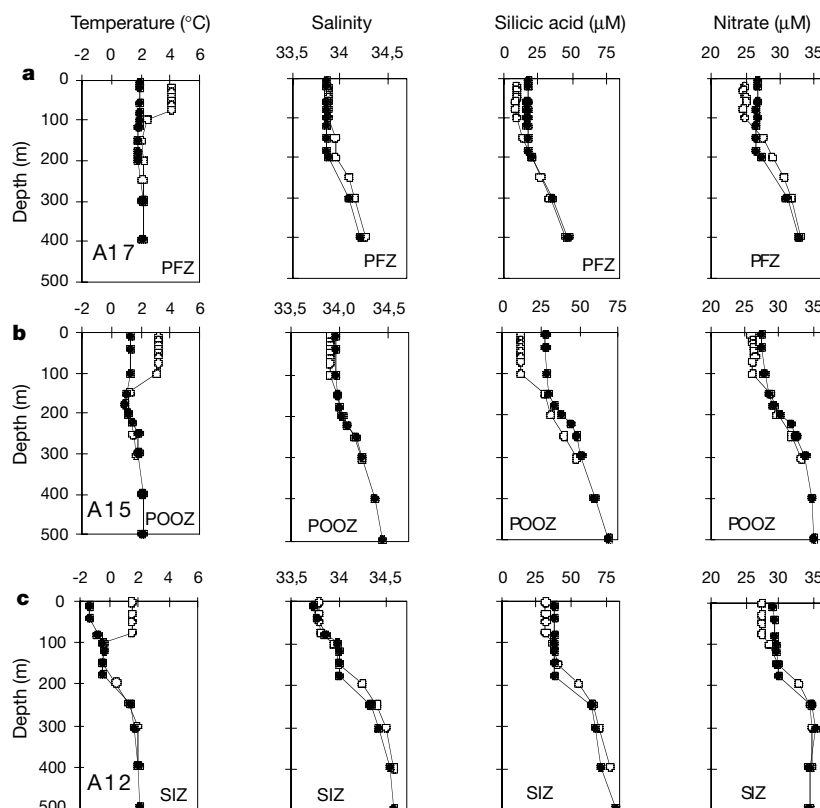


Figure 1 Hydrographic properties at three ANTARES stations. Filled circles, the late winter ANTARES 3 cruise (11–17 October 1994); open circles, the early summer ANTARES 2 cruise (2–7 March 1994). **a**, Station A17 (62°E – $50^\circ 30'\text{S}$) located in the Polar Front zone (PFZ); **b**, station A15 (62°E , $53^\circ 30'\text{S}$) located in the permanently open ocean zone

(POOZ); and **c**, station A12 (62°E , $58^\circ 00'\text{S}$) located at the northern limit of the seasonal ice zone (SIZ). The unaltered remnant winter water observed at the early summer ANTARES 2 stations is located at the depth of the temperature minimum, that is, 100–200 m.

Table 1 Estimated production rates for the Indian sector of the Southern Ocean during spring–summer 1994

Provinces*	Nutrient depletion†		Correction‡		f-ratio§	Estimated production rates over 90 days		
	ΔSi	ΔNO_3^-	C_{Si}	C_{N}		P_{Si}	P_{N}	P_{C}
	(mmol m ⁻²)		(mol m ⁻²)			(mmol m ⁻² d ⁻¹)		
PFZ	1.44 ± 0.37	0.53 ± 0.10	0.38 ± 0.03	0.21 ± 0.05	0.51 ± 0.14	20.2 ± 4.4	16.1 ± 8.8	90.7 ± 49.6
POOZ	2.31 ± 0.49	0.53 ± 0.10	0.42 ± 0.05	0.21 ± 0.05	0.61 ± 0.17	30.3 ± 6.0	13.5 ± 6.7	76.1 ± 37.8
SIZ	0.69 ± 0.26	0.14 ± 0.07	0.51 ± 0.10	0.23 ± 0.06	0.51 ± 0.18	13.3 ± 4.0	8.1 ± 5.6	45.7 ± 31.6

* Provinces: Polar Front zone (PFZ; $<51^\circ\text{S}$), permanently open ocean zone (POOZ; $>51^\circ\text{S}$ and $<58^\circ\text{S}$), and seasonal ice zone (SIZ; $\geq 58^\circ\text{S}$).

† Average vertically integrated nutrient depletion ($\pm$ s.d.) at the Antares 2 stations.

‡ Average correction (C , $\pm$ s.d.) from diffusion (D), plus advection (A), and plus regeneration (R) of nutrients within the surface mixed layer. The province average for D_{Si} and $D_{\text{NO}_3^-}$ are $0.04 \pm 0.01 \text{ mol m}^{-2}$ and $0.01 \pm 0.01 \text{ mol NO}_3^- \text{ m}^{-2}$ for the PFZ, $0.08 \pm 0.03 \text{ mol Si m}^{-2}$ and $0.01 \pm 0.01 \text{ mol NO}_3^- \text{ m}^{-2}$ for the POOZ, and $0.17 \pm 0.08 \text{ mol Si m}^{-2}$ and $0.02 \pm 0.02 \text{ mol NO}_3^- \text{ m}^{-2}$ for the SIZ. The advection and the regeneration fluxes are assumed to be distributed uniformly throughout the region of study; $A_{\text{Si}} = 0.28 \text{ mol Si m}^{-2}$ and $A_{\text{NO}_3^-} = 0.03 \text{ mol N m}^{-2}$, and $R_{\text{Si}} = 0.06 \pm 0.02 \text{ mol Si m}^{-2}$ and $R_{\text{NO}_3^-} = 0.17 \pm 0.04 \text{ mol N m}^{-2}$ (see Methods).

§ Average f-ratio ($\pm$ s.d.) using data from Semeneh *et al.*⁹

|| Province average daily production rates ($\pm$ s.d.), with $P_{\text{Si}} = (\Delta\text{Si} + C_{\text{Si}})10^3/90\text{-days}$ and $P_{\text{N}} = [\Delta\text{NO}_3^- + C_{\text{NO}_3^-}]10^3/f\text{-ratio}/90\text{-days}$. P_{C} is calculated using a C/N elemental molar ratio of 62/11 (ref. 4).

production of biogenic silica reported in Table 1 is 4–5 times higher than the highest measurements of $6.76 \text{ mmol Si m}^{-2} \text{ d}^{-1}$ recorded in the Pacific sector of the POOZ², and 5–8 times higher than the highest measurements of $4.40 \text{ mmol Si m}^{-2} \text{ d}^{-1}$ recorded during the late winter and late summer ANTARES 3 and ANTARES 2 cruises¹³, respectively. Note that the same difference holds for the measured and estimated nitrogen and carbon production rates. As already argued⁴, the method used here for estimating the silicon and nitrogen production rates in the surface layer is conservative, as it inherently records short-time events such as blooms which may have been missed by discrete sampling in the POOZ. Direct measurements of biogenic silica export fluxes recorded by sediment traps deployed in the POOZ at 52°S during austral summer 1994/1995 provide direct evidence for the occurrence of a diatom bloom before the ANTARES 2 cruise (Fig. 3). Fluxes recorded by traps in late February 1994 reach a maximum of $12.66 \text{ mmol Si m}^{-2} \text{ d}^{-1}$ and

of $1.85 \text{ mmol C m}^{-2} \text{ d}^{-1}$ at $1,271 \text{ m}$ (Fig. 3), among the highest values ever recorded for biogenic silica export fluxes.

Spring–summer production rates (Table 1) plus the estimate of the autumn–winter production (see Methods) results in an annual biogenic silica production for the POOZ of $3.34 \pm 0.54 \text{ mol Si m}^{-2} \text{ yr}^{-1}$ (Table 2), being ~ 3 –4 times higher than the upper limit of $1 \text{ mol Si m}^{-2} \text{ yr}^{-1}$ proposed by Nelson *et al.*². Similarly, the estimated annual carbon production for the POOZ of $8.91 \pm 2.79 \text{ mol C m}^{-2} \text{ yr}^{-1}$ is ~ 2 –3 times higher than previous estimates proposed for this area^{5,14}. Note that our estimate of the annual carbon production for the POOZ is consistent with model-derived estimates for the same area ($7.60 \text{ mol C m}^{-2} \text{ yr}^{-1}$, ref. 15), and with recent results obtained from satellite data¹⁶.

Compared with other oceanic basins (Table 2), the estimated annual biogenic silica production rate for the POOZ is 2–3 times higher than in the high-nutrient low chlorophyll (HNLC) equatorial

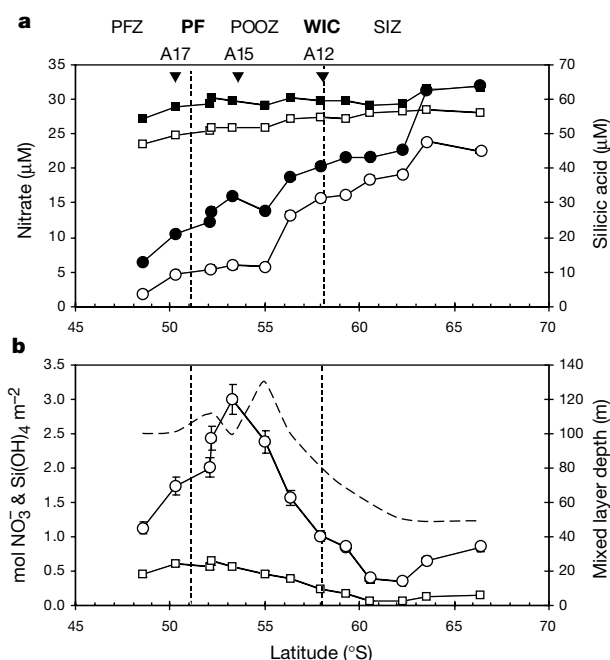


Figure 2 Latitudinal distribution of nutrients and vertically integrated nutrient depletion in the Indian sector of the Southern Ocean during summer 1994. **a**, Concentrations of dissolved silicic acid (circles) and nitrate (squares) in the summer surface water (open symbols) and remnant winter water (filled symbols) of the Indian sector of the Southern Ocean at the early summer ANTARES 2 stations. **b**, Estimated vertically integrated nutrient depletion of silicic acid (circles) and nitrate (squares) at the early summer ANTARES 2

stations ($\pm 7\%$). The dashed line indicates the depth of the mixed layer. The position of the Polar Front (PF) during early summer 1994 is indicated with the dotted line; the northernmost limit of the winter ice cover is also indicated (WIC). These limits defined the three provinces, namely the Polar Front zone (PFZ), the permanently open ocean zone (POOZ) and the seasonal ice zone (SIZ). A17, A15 and A12 are the position of the stations reported in Fig. 1.

Table 2 Annual production of biogenic silica, opal rain rate and ^{230}Th -normalized accumulation rates of opal

Flux ($\text{mol Si m}^{-2} \text{ yr}^{-1}$)	Indian sector of the Southern Ocean*			Equatorial Pacific Ocean		North Atlantic Ocean		Global mean (ref. 3)
	PFZ	POOZ	SIZ	HNLC area†		Oligotroph‡	Mesotroph§	
Production	2.43 ± 0.39	3.34 ± 0.54	1.62 ± 0.58	1.42		0.24	0.30 ± 0.07	0.69 ± 0.12
Rain	—	0.51	0.05	0.14		0.02	0.062 ± 0.025	0.08 ± 0.05
Accumulation	0.075 ± 0.042	0.210 ± 0.040	—	0.016		0.001	0.008	0.017 ± 0.003
WCPE (%)	—	15.3 ± 2.5	3.1 ± 1.1	9.9		8.3	20.7 ± 13.2	11.6 ± 9.2
OPE (%)¶	3.1 ± 2.2	6.3 ± 2.2	—	1.1		0.4	2.7 ± 0.6	2.5 ± 0.9

* This study. Average annual production rates ($P_{\text{Si}} \pm \text{s.d.}$) estimated at the ANTARES 2 transect (1994); $P_{\text{Si}} = \Delta \text{Si} + C_{\text{Si}} + W_{\text{Si}}$, with ΔSi and C_{Si} as reported in Table 1, and with the winter–autumn production (W_{Si}) varying from $0.61 \text{ mol Si m}^{-2}$ within the PFZ and the POOZ to $0.42 \pm 0.22 \text{ mol Si m}^{-2}$ within the SIZ (see Methods). Annual rain rates measured by means of sediment traps (depth $\geq 3,000 \text{ m}$) between February 1994 and February 1995; POOZ: 52°S – 61°E ; SIZ: 63°S – 70°E (ANTARES programme). ^{230}Th -normalized accumulation rates ($\pm \text{s.d.}$) in the same region; PFZ: cores K1P11 (49°S – 58°E) and K1B13 (50°S – 58°E); POOZ: cores K1B06 (52°S – 61°E), MD84-551 (55°S – 73°E) and MD84-552 (55°S – 71°E).

† Equatorial Pacific; references: production (ref. 17); rain rate (ref. 20); accumulation (ref. 26).

‡ Western Sargasso Sea; references: production (ref. 18); rain rate (ref. 21); accumulation (ref. 27).

§ Northeast Atlantic; references: production (ref. 15 and O.R., unpublished data); rain rate and accumulation (O.R., unpublished data).

|| WCPE: water column preservation efficiency (rain/production).

¶ OPE: overall preservation efficiency (accumulation/production). Note that the historical estimates for production and accumulation rates in the PFZ and POOZ were $1.0 \text{ mol Si m}^{-2} \text{ yr}^{-1}$ (ref. 2) and 0.40 – $0.42 \text{ mol Si m}^{-2} \text{ yr}^{-1}$ (without correction from focusing; ref. 25), respectively; that is, an OPE of $\sim 40\%$ for this area.

Pacific¹⁷, and more than one order of magnitude higher than in the Atlantic Ocean (refs 15 and 18, and O.R., unpublished data). Similarly, water-column annual export fluxes in the POOZ of $0.51 \text{ mol Si m}^{-2} \text{ yr}^{-1}$ (Table 2) are the highest reported¹⁹. They are 3–4 times higher than in the equatorial Pacific²⁰ and 4–25 times higher than in the North Atlantic (ref. 21 and O.R., unpublished data). Combining production and sediment-trap data results in an estimate of water-column preservation efficiency of $15.3 \pm 2.3\%$ for the POOZ (Table 2). This value is 23–53% higher than the value for the oligotrophic Atlantic or the HNLC equatorial Pacific, but similar to the value for the mesotrophic northeast Atlantic (Table 2). We note that the areas with the highest opal water-column preservation efficiency are sites with a high seasonality²², and exhibit the highest export/production ratio for carbon²².

These results have implications in the context of the ‘opal paradox’, by which high opal accumulation would occur in sediments despite low production of biogenic silica in overlying surface waters². Nelson *et al.*² suggested an overall preservation efficiency substantially higher than anywhere else as an explanation for this paradox. However, the calculations we report above can be combined with opal accumulation rates to calculate overall preservation efficiencies. Note that opal accumulation rates may be affected by post-depositional sediment redistribution by bottom currents. This problem has been alleviated here by normalizing opal flux to excess ²³⁰Th activity in sediments^{23,24}, a method that corrects for sediment focusing or winnowing and provides estimates of preserved accumulation rate. Thus, an opal accumulation rate of $0.210 \pm 0.040 \text{ mol Si m}^{-2} \text{ yr}^{-1}$ can be estimated for sediments underlying the Indian sector of the POOZ; this rate is ~2 times lower than previously thought²⁵ (Table 2). Although focusing has been corrected for, this value is respectively 12, 31 and 185 times higher than for the equatorial Pacific²⁶, the mesotrophic northeast Atlantic (O.R., unpublished data) and the oligotrophic Atlantic²⁷ (Table 2). The overall preservation efficiency of $6.3 \pm 2.2\%$ inferred for the POOZ is thus 5–10 times lower than the value of 40% previously thought (Table 2). A similar observation is valid for the

PFZ with an overall preservation efficiency of $3.1 \pm 2.2\%$, that is, 7–45 times lower than previously thought (Table 2). This calculation shows that the overall preservation efficiency for the Indian sector of the POOZ is only ~2 times higher than the global mean of $2.5 \pm 0.9\%$ (ref. 3); it is also only ~2 times higher than, for example, the value for the northeast Atlantic where little opal accumulates (Table 2). For the PFZ, the overall preservation efficiency is similar to the global mean and to the value for the northeast Atlantic.

This calculation shows that opal accumulation is higher in the Southern Ocean than anywhere else; this is mostly because of a higher biogenic opal rain rate, in turn due to higher biogenic silica production rates in surface waters. Spatial differences in opal preservation play a role, but as a modulation factor only and not as a primary factor. This conclusion has implications for the use of the opal sedimentary record as a proxy for past export siliceous productivity. If opal is to be used as a tracer of palaeoproductivity (and of the efficiency of the biological pump) in response to past climate change, we need to achieve a better understanding of the factors controlling the degree of coupling between the silicon and carbon biogeochemical cycles—both in the water column and in the sediments. □

Methods

Biogenic silica daily production rates

Estimates of biogenic silica daily production rates are derived from integrated silicic acid depletion (ΔSi) by taking into account, during the 90 d of the productive period, the input of silicic acid into the surface layer by (1) vertical diffusion (D_{Si}), (2) lateral advection (A_{Si}), and (3) the regeneration of silicic acid through dissolution of biogenic silica within the surface mixed layer (R_{Si}). Thus, the spring–summer production of biogenic silica (P_{Si}) is written:

$$P_{\text{Si}} = \Delta\text{Si} + D_{\text{Si}} + A_{\text{Si}} + R_{\text{Si}} \quad (1)$$

The diffusive flux of silicic acid through the pycnocline during spring–summer, D_{Si} , is calculated using the diffusion coefficient $K_z = 0.5 \text{ cm}^2 \text{ s}^{-1}$ (ref. 28), and the vertical gradient of silicic acid at the ANTARES 2 stations. From north to south, this gradient ranges between 0.09 to 0.66 mmol Si m^{-4} , with maxima occurring within the SIZ. Thus, the calculated diffusion rates varies from 0.33 within the PFZ to 2.89 $\text{mmol Si m}^{-2} \text{ d}^{-1}$ within the SIZ, or 0.03–0.26 mol Si m^{-2} integrated over a period of 90 d. The advective flux, A_{Si} , is calculated assuming a permanent input of silicic acid at the Antarctic Divergence (AD), and a permanent output of silicic acid through the Antarctic Intermediate Water (AAIW) at the Polar Front (PF), following the calculation proposed by Jacques⁵. Thus, we assume that A_{Si} is distributed uniformly over the transect. The concentration of silicic acid in the flow entering the surface layer at the AD is ~60 mmol m^{-3} ; in the AAIW it has dropped down to ~15 mmol m^{-3} . With a surface area of $26 \times 10^{12} \text{ m}^2$ between AD and PFZ, the northward current being $18 \times 10^{11} \text{ m}^3 \text{ d}^{-1}$, we infer a mean silicic acid advective flux of $3.11 \text{ mmol Si m}^{-2} \text{ d}^{-1}$ (that is, 0.28 mol Si m^{-2} per 90 d). Although a uniform distribution of A_{Si} is assumed, the error introduced may be small compared to the value of ΔSi (Table 1).

The regeneration term R_{Si} is estimated from the latitudinal-average biogenic silica concentration within the mixed layer during summer— $70.92 \pm 26.20 \text{ mmol Si m}^{-2}$ along the ANTARES 2 transect¹³—assuming the specific dissolution rate of biogenic silica to be 0.010 d^{-1} for this region²⁹. Thus, the dissolution rate within the mixed layer for the Indian sector is estimated to be $0.71 \pm 0.26 \text{ mmol Si m}^{-2} \text{ d}^{-1}$, that is, R_{Si} equals $0.06 \pm 0.02 \text{ mol Si m}^{-2}$ per 90 d. Note that R_{Si} is assumed to be uniform throughout the studied area. Although this may not be strictly true, the numerical value of R_{Si} is small compared to the other terms (Table 1).

Nitrogen daily production rates

Estimates of nitrogen production rates are derived from vertically integrated nitrate depletion by taking into account the same correcting factors as for biogenic silica. Thus, the spring–summer production of particulate organic nitrogen (P_{N}) is written:

$$P_{\text{N}} = [\Delta\text{NO}_3^- + D_{\text{NO}_3} + A_{\text{NO}_3} + R_{\text{NO}_3}] / f\text{-ratio} \quad (2)$$

From north to south, the nitrate gradient ranges from 0.02 to 0.14 $\text{mmol NO}_3^- \text{ m}^{-4}$. Thus, the diffusive flux of nitrate, D_{NO_3} , varies from 0.11 to 0.56 $\text{mmol NO}_3^- \text{ m}^{-2} \text{ d}^{-1}$ (equivalent to 0.01–0.05 $\text{mol NO}_3^- \text{ m}^{-2}$ per 90 d). In addition, the advective flux, A_{NO_3} , is estimated to be 0.28 $\text{mmol NO}_3^- \text{ m}^{-2} \text{ d}^{-1}$ (ref. 5) (equivalent to 0.03 mol N m^{-2} per 90 d).

For the late summer ANTARES1 stations, the average nitrite oxidation rate is $1.92 \pm 0.42 \text{ mmol N m}^{-2} \text{ d}^{-1}$ (ref. 30); that is, R_{NO_3} equals $0.17 \pm 0.04 \text{ mol N m}^{-2}$ per 90 d.

In addition, in equation (2), the fraction of total nitrogen production sustained by ammonium or urea is deduced from shipboard determination of the f -ratio at the ANTARES 2 stations⁹ (Table 1).

Estimate of the annual Si and C production

The calculation of the annual Si and C production (P_{Si}^{a} and P_{C}^{a}) requires the estimation of

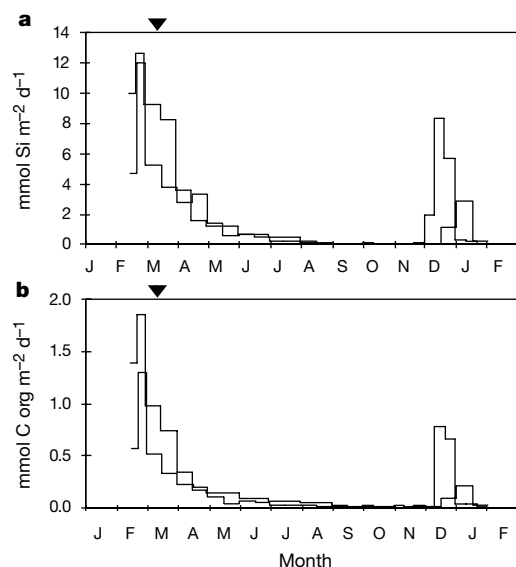


Figure 3 Average daily export fluxes of particulate organic Si and C in the Indian sector of the Southern Ocean. **a**, Data for biogenic silica, and **b**, data for organic carbon, measured by sediment traps at 52°S–62°E in the Indian sector of the Southern Ocean (POOZ) between January 1994 and January 1995 (ANTARES Southern Ocean-JGOFs programme; P.T., unpublished data). The sediment traps were positioned at depths of 1,271 m (bold line) and 4,023 m (thin line). The arrows represent the period of the ANTARES 2 cruise (2–7 March 1994).

the part of the production that takes place during autumn and winter (W_{Si} and W_C , which encompass 275 d). Thus, the annual production of Si and C are written:

$$P_{Si}^* = P_{Si} + W_{Si} \quad (3a)$$

$$P_C^* = P_N(C/N) + W_C \quad (3b)$$

In equation (3b), $C/N = 62/11$ (mol mol⁻¹) is the cellular elemental composition for this area¹. The autumn–winter productivity is estimated from discrete measurements of Si and C production rates, using the ³²Si and ¹⁴C isotopic technique, during the late winter ANTARES 3 cruise¹³. In the ice-free area, the productivity was 2.2 mmol Si m⁻² d⁻¹, and 7.5 mmol C m⁻² d⁻¹. Assuming constancy throughout winter and autumn in the ice-free area, this leads to a maximum of $W_{Si} = 0.61$ mol Si m⁻² and $W_C = 2.06$ mol C m⁻². Note that these values are consistent, within $\pm 15\%$, with model-derived estimates of winter and autumn production for the ice-free Indian sector of the Southern Ocean¹⁵. South of 58° S, that is, the northernmost limit of the ice extent during winter, W_{Si} and W_C were corrected using estimates of the seasonal ice-cover from the ERS1 satellite. For each month, we assume that the production is negligible when a station is totally covered by sea ice.

Combining Tables 1 and 2 shows that the spring–summer production represents from 70% (PFZ) to >90% (SIZ) of the annual biogenic silica production.

Received 22 October; accepted 1 March 2000.

- Boyle, E. A. J. Pumping iron makes thinner diatoms. *Nature* **393**, 733–734 (1998).
- Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A. & Quéguiner, B. Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Glob. Biogeochem. Cycles* **9**, 359–372 (1995).
- Tréguer, P. et al. The silica balance in the world ocean—a reestimate. *Science* **268**, 375–379 (1995).
- Jennings, J. C. Jr, Gordon, L. I. & Nelson, D. M. Nutrient depletion indicates high primary productivity in the Weddell Sea. *Nature* **309**, 51–54 (1984).
- Jacques, G. Is the concept of new-regenerated production valid for the Southern Ocean? *Mar. Chem.* **35**, 273–286 (1991).
- Minas, H. J. & Minas, M. Net community production in “High Nutrient-Low Chlorophyll” waters of the tropical and Antarctic Ocean: grazing vs iron hypothesis. *Oceanol. Acta* **15**, 145–162 (1992).
- Clowes, A. J. Phosphate and silicate in the Southern Ocean. *Discovery Rep.* **XIX**, 1–120 (1938).
- Kamykowski, D. & Zentara, S. J. Circumpolar plant nutrient covariation in the Southern Ocean: pattern and processes. *Mar. Ecol. Prog. Ser.* **58**, 101–111 (1989).
- Semench, M., Dehairs, F., Fiala, M., Elskens, M. & Goeys, L. Seasonal variation of phytoplankton community structure and nitrogen uptake regime in the Indian sector of the Southern Ocean. *Polar Biol.* **20**, 259–272 (1998).
- Davis, C. O. Continuous culture of marine diatoms under silicate limitation. II. Effect of light intensity on growth and nutrient uptake of *Skeletonema costatum*. *J. Phycol.* **12**, 291–300 (1976).
- Takeda, S. Influence of iron availability on nutrient consumption ratio of diatoms in oceanic waters. *Nature* **393**, 774–777 (1998).
- Quéguiner, B., Tréguer, P., Pecken, I. & Shareck, R. Biogeochemical dynamics and the silicon cycle in the Atlantic sector of the Southern Ocean during Austral spring 1992. *Deep-Sea Res.* **44**, 68–89 (1997).
- Caubert, T. Couplage et découplage des cycles du carbone et du silicium dans le secteur Indien de l’Océan Austral. Thesis, Univ. Bretagne Occidentale (1998).
- Wefer, G. & Fischer, G. Annual primary production and export flux in the Southern Ocean from sediment trap data. *Mar. Chem.* **35**, 597–613 (1991).
- Pondaven, P., Ruiz-Pino, D., Druon, J. N., Fravallo, C. & Tréguer, P. Factors controlling silicon and nitrogen biogeochemical cycles in high nutrient, low chlorophyll systems (the Southern Ocean and the North Pacific – Comparison with a mesotrophic system (the North Atlantic). *Deep-Sea Res.* **46**, 1923–1968 (1999).
- Arrigo, K., Worthon, D., Schnell, A. & Lizotte, M. P. Primary production in Southern Ocean waters. *J. Geophys. Res.* **103**, 15587–15600 (1998).
- Blain, S., Leynaert, A., Tréguer, P., Chrétiennot-Dinet, M.-J. & Rodier, M. Biomass, growth rates and limitation of equatorial Pacific diatoms. *Deep-Sea Res.* **44**, 1255–1275 (1997).
- Nelson, D. M. & Brzezinski, M. A. Diatom growth and productivity in an oligotrophic midocean gyre: a 3-yr record from the Sargasso Sea near Bermuda. *Limnol. Oceanogr.* **42**, 473–486 (1997).
- Lampitt, R. & Antia, A. N. Particle flux in deep-seas: regional characteristics and temporal variability. *Deep-Sea Res.* **44**, 1377–1403 (1997).
- Honjo, S., Dymond, J., Collier, R. & Manganini, S. J. Export production of particles to the interior of the Equatorial Pacific Ocean during the 1992 Eqpac experiment. *Deep-Sea Res.* **42**, 831–870 (1995).
- Deuser, W. G., Jickells, T. D., King, P. & Commeau, J. A. Decadal and annual changes in biogenic opal and carbonate fluxes to the deep Sargasso Sea. *Deep-Sea Res.* **33**, 225–246 (1995).
- Berger, W. H. & Wefer, G. Export production: seasonality and intermittency, and paleoceanographic implications. *Paleoceanogr. Paleoclimatol. Paleocool.* **89**, 245–254 (1990).
- Bacon, M. P. Glacial to Inter-glacial changes in carbonate and clay sedimentation in the Atlantic ocean estimated from ²³⁰Th measurements. *Isotope Geosci.* **2**, 97–111 (1984).
- Francois, R., Bacon, M. P., Altabet, M. A. & Labeyrie, L. D. Glacial/interglacial changes in sediment rain rate in the SW Indian sector of Subantarctic waters as recorded by ²³⁰Th, ²³¹Pa, U, and ¹⁵N. *Paleoceanography* **8**, 611–629 (1993).
- Rabouille, C., Gaillard, J. F., Tréguer, P. & Vincendeau, M. A. Biogenic silica recycling in surficial sediments across the Polar Front of the Southern Ocean (Indian Sector). *Deep-Sea Res.* **44**, 1151–1176 (1997).
- Berelson, W. M. et al. Biogenic budgets of particle rain, benthic remineralization and sediment accumulation in the Equatorial Pacific. *Deep-Sea Res.* **44**, 2251–2282 (1997).
- Sayles, F. L. et al. The benthic cycle of biogenic opal at the Bermuda Atlantic Time series site. *Deep-Sea Res.* **43**, 383–409 (1996).
- Gordon, A. L., Chen, C. T. A. & Metcalf, W. G. Winter layer entrainment of the Weddell Deep Water. *J. Geophys. Res.* **89**, 637–640 (1984).
- Tréguer, P., Kamatani, A., Gueñe, S. & Quéguiner, B. Kinetics of dissolution of Antarctic diatoms frustules and the biogeochemical cycle of silicon in the Southern Ocean. *Polar Biol.* **9**, 397–403 (1989).
- Bianchi, M., Feliatra, F., Tréguer, P., Vincendeau, M. & Morvan, J. Nitrification rates, ammonium and nitrate distribution in upper layers of the water column and in sediments of the Indian sector of the Southern Ocean. *Deep-Sea Res.* **44**, 1017–1032 (1997).

Acknowledgements

We thank J. Morvan, C. Jeandel, M. Fiala and P. Mayzaud for taking care of the moorings and sediment traps during the ANTARES programme; the officers, engineers and crew of RV *Marion Dufresne* for their help during the ANTARES cruises; CERSAT (IFREMER/Plouzané, France) for providing ERS1 satellite estimates of the seasonal ice-cover; R. Lampitt and A. Gooday for taking care of the moorings, sediment traps and multiple cores during the BENGAL programme in the North East Atlantic; and R. C. Dugdale for critically reading the manuscript. This work was supported by Région Bretagne (P.P.), the Institut National des Sciences de l’Univers (INSU/CNRS), and the French Polar Research Institute (IFRTP/CNRS).

Correspondence and requests for materials should be addressed to P.P. (e-mail: philippe.pondaven@univ-brest.fr).

Constraints on the composition of the Earth’s core from *ab initio* calculations

D. Alfé*, M. J. Gillan† & G. D. Price*

* Research School of Geological and Geophysical Sciences, Birkbeck College and University College London, Gower Street, London WC1E 6BT, UK

† Physics and Astronomy Department, University College London, Gower Street, London WC1E 6BT, UK

Knowledge of the composition of the Earth’s core^{1–3} is important for understanding its melting point and therefore the temperature at the inner-core boundary and the temperature profile of the core and mantle. In addition, the partitioning of light elements between solid and liquid, as the outer core freezes at the inner-core boundary, is believed to drive compositional convection⁴, which in turn generates the Earth’s magnetic field. It is generally accepted that the liquid outer core and the solid inner core consist mainly of iron¹. The outer core, however, is also thought to contain a significant fraction of light elements, because its density—as deduced from seismological data and other measurements—is 6–10 per cent less than that estimated for pure liquid iron^{1–3}. Similar evidence indicates a smaller but still appreciable fraction of light elements in the inner core^{5,6}. The leading candidates for the light elements present in the core are sulphur, oxygen and silicon⁷. Here we obtain a constraint on core composition derived from *ab initio* calculation of the chemical potentials of light elements dissolved in solid and liquid iron. We present results for the case of sulphur, which provide strong evidence against the proposal that the outer core is close to being a binary iron–sulphur mixture⁷.

Our proposed constraint is simply stated. If, as commonly assumed, the solid and liquid are in thermodynamic equilibrium at the inner-core boundary (ICB), the chemical potentials of a given element in both phases must be equal at the boundary. This condition determines the ratio of concentrations of elements in the solid and the liquid phase, and hence the relation between the outer-core and inner-core densities, which must agree with seismic measurements of these densities.

The chemical potential μ_X of a solute X in a solid or liquid solution is conventionally expressed as $\mu_X = \mu_X^\dagger + k_B T \ln a_X$, where $\mu_X^\dagger$ is a constant and a_X is the activity⁸. To reflect the fact that a_X becomes equal to the mole fraction c_X in the dilute limit $c_X \rightarrow 0$, it is common practice to write $a_X = \gamma_X c_X$, where the activity coefficient, γ_X , has the property $\gamma_X \rightarrow 1$ as $c_X \rightarrow 0$ (ref. 8). Here we rearrange the expression for μ_X as $\mu_X = \tilde{\mu}_X + k_B T \ln c_X$, where $\tilde{\mu}_X$ is defined to be equal to $\mu_X^\dagger + k_B T \ln \gamma_X$. Equality of the chemical potentials μ_X and

μ_X^s in coexisting liquid and solid (superscripts l and s respectively) then requires that

$$\tilde{\mu}_X^l + k_B T \ln c_X^l = \tilde{\mu}_X^s + k_B T \ln c_X^s \quad (1)$$

or equivalently

$$c_X^l/c_X^s = \exp[(\tilde{\mu}_X^l - \tilde{\mu}_X^s)/k_B T] \quad (2)$$

This means that the ratio of the mole fractions c_X^l and c_X^s of X in the solid and liquid solution is determined by the liquid and solid thermodynamic quantities $\tilde{\mu}_X^l$ and $\tilde{\mu}_X^s$. Although liquid–solid equilibrium in the iron–sulphur (Fe–S) system has been experimentally studied up to pressures of around 60 GPa (ref. 9), there seems little immediate prospect of obtaining experimental data for $\tilde{\mu}_X^l - \tilde{\mu}_X^s$ for Fe–S or any other binary Fe– X system at the much higher ICB pressure of 330 GPa. However, we will show here that a fully *ab initio* calculation of $\tilde{\mu}_X^l$ and $\tilde{\mu}_X^s$ is achievable for S.

If we can obtain accurate values of $\tilde{\mu}_X^l$ and $\tilde{\mu}_X^s$, then we can test any binary alloy (Fe– X) model for the core as follows: (1) we use *ab initio* calculations on the iron alloy liquid at the ICB pressure to determine the liquid mole fraction c_X^l needed to reproduce the measured ICB density of the outer core; (2) we use equation (2) to determine c_X^s ; (3) we use *ab initio* calculations on the Fe– X solid with this c_X^s value to deduce its density at the ICB pressure. If this approach disagrees with the density deduced from seismic measurements¹⁰ the model can be ruled out. (Allowance must, of course, be made for uncertainties in both seismic and *ab initio* values.)

The use of *ab initio* molecular dynamics based on density-functional theory (DFT) to investigate liquid Fe– X alloys under Earth's core conditions, and in particular to determine the impurity mole fractions needed to account for a specified density reduction, has been described elsewhere. In our earlier work on liquid Fe–S (refs 11, 12), we showed that at ICB pressures an S mole fraction of $c_S^l = 18.75\%$ reduces the density from that of pure Fe by 8%. In fact, our most recent calculations¹³ show that the density of pure liquid Fe at ICB conditions is $\sim 6\%$ higher than the seismic density, so that the required mole fraction of S is estimated to be equal to 15%. This is close to estimates in the region of 16% deduced from experimental equation-of-state measurements¹⁴.

The techniques developed here for calculating chemical potentials build on recent progress in *ab initio* calculations of free energies^{15–18}. We start from the basic definition that μ_X (per atom) can either represent the change of Gibbs free energy when an atom of X is added to the system at constant temperature T and pressure P , or the change of Helmholtz free energy F at constant T and volume V . Computationally, it is convenient to work with the difference $\mu_X - \mu_{Fe}$ of chemical potentials of X and Fe, which represents the change of free energy when an atom of Fe is replaced by an atom of X . The chemical potential μ_X and hence $\tilde{\mu}_X$ is obtained from $\mu_X - \mu_{Fe}$ by making use of the Gibbs free energies of pure solid and liquid Fe calculated in our recent work¹⁶. (In the following, where necessary, we shall use the quantity $\tilde{\mu}_{Fe}$ defined in the analogous way to $\tilde{\mu}_X$ by the equation $\mu_{Fe} = \tilde{\mu}_{Fe} + k_B T \ln c_{Fe}$, where $c_{Fe} = 1 - c_X$ represents the mole fraction of Fe.)

For the crystal, the change of F associated with formation of a substitutional X impurity by replacement of an Fe atom by an X atom is readily computed in the limit $c_X \rightarrow 0$. (For S, O and Si impurities, interstitial formation is expected to be energetically very unfavourable, and can be ignored.) In the expression $F - E - TS$, the change of internal energy E is the difference between the *ab initio* energy of the perfect Fe crystal and that of the same crystal with a single substitutional X atom. The contribution from the change of entropy S can be expressed in the harmonic approximation as $-k_B T \sum_n [\ln(\hbar\omega_n'/k_B T) - \ln(\hbar\omega_n/k_B T)]$, where ω_n and ω_n' are the lattice vibrational frequencies of the Fe crystal without and with the X atom. The *ab initio* calculation of lattice vibrational frequencies is well-established, and presents no major problems¹⁹. Here we use our

implementation of the small-displacement method^{17,20} to calculate all the ω_n and ω_n' in a large supercell of crystal. We will return below to the estimation of anharmonic corrections.

To calculate μ_X^s away from the dilute limit, we note that the part of the statistical mechanics associated with the rearrangement of X atoms on lattice sites is rigorously equivalent to a lattice-gas problem, for which we can use standard Monte Carlo methods²¹ to evaluate the free energy. The crucial input from *ab initio* calculations is the interaction free energy of X atoms when they are on neighbouring sites. This interaction free energy is just the change in F when a pair of X atoms, initially on distant sites, are placed on nearest-neighbour sites. The change in E in this process is obtained from a straightforward *ab initio* calculation, and the contribution from the entropy change is also obtained from an *ab initio* calculation of the lattice vibrational frequencies for the X – X pair on neighbouring sites.

For liquid Fe– X , we calculated the difference $\mu_X - \mu_{Fe}$ using ‘thermodynamic integration’²². This is a general technique for computing the Helmholtz free energy difference $F_1 - F_0$ of two systems containing the same number N of atoms, but having different total energy functions $U_1(\mathbf{r}_1, \dots, \mathbf{r}_N)$ and $U_0(\mathbf{r}_1, \dots, \mathbf{r}_N)$, where $\mathbf{r}_i$ is the position of atom i . The difference $F_1 - F_0$ represents the reversible work done when continuously switching the total energy function from U_0 to U_1 at constant volume, and can be expressed as

$$F_1 - F_0 = \int_0^1 d\lambda \langle U_1 - U_0 \rangle_\lambda \quad (3)$$

where the average $\langle \dots \rangle$ is calculated in thermal equilibrium for the system governed by the ‘hybrid’ energy function $U_\lambda = (1 - \lambda)U_0 + \lambda U_1$. This well-established *ab initio* calculation technique of liquid-state free energies^{15,23} was used in our recent investigation¹⁶ of the high-pressure melting curve of Fe. In order to compute $\mu_X - \mu_{Fe}$, we let U_0 be the total energy for the system with N_{Fe} atoms of Fe and N_X of X , and we let U_1 be the same for $N_{Fe} - 1$ atoms of Fe and $N_X + 1$ of X . We evaluate $\langle U_1 - U_0 \rangle_\lambda$ by performing *ab initio* molecular dynamics with time evolution generated by U_λ , and taking the time average of $U_1 - U_0$. We repeat this procedure for several values of λ , and the integration over λ is done numerically.

This type of ‘transmutation’ of Fe into S obviously does not correspond to a real-world process, but in terms of *ab initio* statistical mechanics is a perfectly rigorous way of obtaining the chemical potential difference $\tilde{\mu}_X - \tilde{\mu}_{Fe}$. It demands an unusual kind of simulation: for the atom positions $\mathbf{r}_1, \dots, \mathbf{r}_N$ at each instant of time, we have to perform two independent *ab initio* calculations, one for each chemical composition. As well as U_0 and U_1 for the given positions, we calculate two sets of *ab initio* forces $\mathbf{F}_{0i} = -\nabla_i U_0$ and $\mathbf{F}_{1i} = -\nabla_i U_1$, and the linear combinations $\mathbf{F}_{\lambda i} = (1 - \lambda)\mathbf{F}_{0i} + \lambda\mathbf{F}_{1i}$ are used to generate the time evolution. In practice, the statistical accuracy is rather poor if one ‘transmutes’ only a single Fe atom into X , and it is preferable to ‘transmute’ several atoms at the same time. Instead of $\tilde{\mu}_X - \tilde{\mu}_{Fe}$ for a given mole fraction c_X , this then yields an integral of $\mu_X - \mu_{Fe}$ over a range of c_X values. The results obtained by ‘transmuting’ different numbers of atoms can then be processed to obtain $\tilde{\mu}_X - \tilde{\mu}_{Fe}$ as a function of c_X . Once we obtain $\tilde{\mu}_X - \tilde{\mu}_{Fe}$ for both solid and liquid, the difference $\tilde{\mu}_X^l - \tilde{\mu}_X^s$ which determines c_X^l/c_X^s is obtained by using values for μ_{Fe} in pure Fe obtained in our earlier *ab initio* calculations¹⁶.

Our practical *ab initio* calculations on the binary Fe–S system employed exactly the same DFT techniques with plane-wave basis sets and ultrasoft pseudopotentials described in detail in our earlier work on the high-pressure Fe–S and Fe–O liquids^{11,12}. The calculations were done using the software package VASP²⁴, which is exceptionally efficient and stable for metallic systems. We have already reported extensive comparisons with a range of experimental data and with other implementations of DFT which demon-

strate the accuracy and reliability of the techniques¹³. We focus mainly on results for the S chemical potentials obtained at the thermodynamic state $P = 370$ GPa and $T = 7,000$ K, which is close to the ICB pressure but at a somewhat higher temperature; results obtained at $P = 80$ GPa and $T = 3,500$ K are mentioned later. Our solid phase is taken to have the hexagonal-close-packed (h.c.p.) structure, because the available evidence indicates that this is the stable structure for Fe under inner-core conditions²⁵.

In the high- P /high- T state, we find the dilute-limit value $\mu_S^{\text{li}} - \mu_S^{\text{so}} = -0.25$ eV. As expected, its negative values favours partitioning of S into the liquid, but its magnitude is smaller than $k_B T$, so that the partitioning is weak: if we put this value into equation (2), we obtain $c_S^{\text{li}}/c_S^{\text{so}} = 0.66$. Even more importantly, we find that the variation of activity coefficients (γ_S) with concentration cannot be ignored, and that both $\tilde{\mu}_S^{\text{li}}$ and $\tilde{\mu}_S^{\text{so}}$ increase strongly with increasing mole fraction of S. For concentrations up to $\sim 20\%$, $\tilde{\mu}_S$ is well represented as $\tilde{\mu}_S = \mu_S^{\text{li}} + \lambda_S c_S$, with respective liquid and solid values of λ_S of 6.3 and 6.1 eV. Equation (2) shows that the similarity of these λ_S values will prevent the solid and liquid concentrations differing significantly. For a value of $c_S^{\text{li}} = 0.15$, we find the solution $c_S^{\text{so}} = 0.13$. The conclusion is that if the outer core is a binary Fe–S mixture, the mole fraction of S in the inner core will be $\sim 13\%$.

This is completely incompatible with seismic measurements, which give an accurate value for the density difference between liquid and solid at the ICB equal to $4.5 \pm 0.5\%$ (ref. 10). According to our earlier *ab initio* calculations¹⁶, incompatibility arises due to a density difference between coexisting liquid and solid pure Fe at the ICB pressure of only 1.6%. Our present calculations show that the S atomic volume in both solid and liquid is almost identical to that of Fe. To a sufficient accuracy it is true that the fractional change of density in each phase can be calculated simply as $c_S(m_S - m_{\text{Fe}})/m_{\text{Fe}}$, where m_S and m_{Fe} are the atomic masses of S and Fe. This means that if c_S^{li} and c_S^{so} were equal, the density difference between liquid and solid would remain unchanged at 1.6%. The small difference between c_S^{li} and c_S^{so} has the effect of increasing the density difference only to 2.5%, which is still well below the seismic value of $4.5 \pm 0.5\%$. We therefore rule out the binary Fe–S model for core composition. Even without density data, we observe that the very high solubility of S in the solid can cause another major problem. It is generally believed that convection in the outer core is driven by the partitioning of light impurities into the liquid, which is also incompatible with the very weak partitioning found here.

If our *ab initio* techniques are reliable, they must yield predictions that are compatible with the known Fe–S phase diagram at low pressures and temperatures. Measurements²⁶ show that Fe–S is a normal eutectic system at ambient pressure, with a strong depression of freezing point, and almost complete partitioning of S into the liquid phase on the Fe-rich side of the phase diagram. Our *ab initio* results at $P = 80$ GPa and $T = 3,500$ K give the dilute-limit value $\mu_S^{\text{li}} - \mu_S^{\text{so}} = -0.36$ eV, which is significantly less than our high- P /high- T value. We also find that the variation of activity coefficients γ_S with concentration is much weaker, with λ_S values equal to 1.4 and 2.4 in the liquid and solid, respectively. At temperatures equal to and below the ambient-pressure melting temperature of Fe (1,810 K), we therefore predict strong partitioning of S to the liquid phase, as expected from the phase diagram.

Although the consequences of our calculated chemical potentials are striking, their values are not. As stressed above, S and Fe atoms are almost exactly the same size. The (free) energetic effects of replacing Fe by S should therefore be almost identical in both phases. Near the Fe melting curve, a difference $\tilde{\mu}_S^{\text{li}} - \tilde{\mu}_S^{\text{so}}$ of more than a few tenths of an eV would therefore have been surprising. The cause of the strong increase of $\tilde{\mu}_S$ with concentration is also very clear. We showed earlier that there is an effective repulsion between S atoms in the high-pressure liquid phase¹¹. This arises because the S valence states lie at least 10 eV below the Fermi energy, so that

chemical bonding between S atoms is very weak. If two isolated S atoms are brought together, two Fe–S bonds are lost and one Fe–Fe bond is gained. The relative strengths of the Fe–S and Fe–Fe bonds mean that this process is energetically unfavourable. We also note that Boehler's measurements⁹ on eutectic behaviour in the Fe–S system up to pressures of ~ 60 GPa indicate that increasing pressure reduces the depression of the freezing point. These measurements were interpreted by suggesting that "Fe and FeS exhibit binary solid-solution behaviour rather than eutectic behaviour at core pressures"⁹, which is exactly what our present calculations show.

Are the *ab initio* calculations precise enough to support our claim that the binary Fe–S model for the core is untenable? We stress that high precision is both expected and unnecessary. DFT calculations are at their best when atoms change neither the electronic structure nor their environment. As the close-packed structures of the solid and liquid are almost identical at ICB pressures, high precision is expected¹³. The obvious omission in our calculations is to neglect anharmonicity in the solid. In our recent calculations of the free energy of h.c.p. Fe under Earth's core conditions^{16,17}, we have shown that the anharmonic contribution to the free energy can be ~ 60 meV per atom on the melting curve, so the error in our values for $\tilde{\mu}_S^{\text{li}} - \tilde{\mu}_S^{\text{so}}$ might be as large as this. But even in the unlikely event that the chemical potential differences varied by several tenths of an eV, our conclusions would not change, as the effective repulsion between S atoms forces near-equalization of the solid and liquid concentrations.

Even if one adopts an extremely cautious view of the precision achievable by present *ab initio* calculations, we believe that technical developments now underway will answer all doubts on this score. Quantum Monte Carlo (QMC) techniques²⁷ have been demonstrated to give considerably higher precision for energetics than the best DFT calculations²⁸, and QMC calculations on condensed-matter systems of well over 100 atoms have recently been reported²⁹. The use of QMC in the near future to tackle systems as challenging as liquid Fe–X alloys is not unthinkable.

The main conclusion of this work is that a simple binary Fe–S core composition is incompatible with the geophysical facts. But the general constraint provided by *ab initio* chemical potentials should be capable of telling us much more than just this. For example, if we take a ternary model, such as Fe–S–O or Fe–S–Si, the constraint should uniquely fix the two impurity concentrations. Recent experiments³⁰ suggest that Si dissolves in liquid Fe only under strongly reducing conditions, so that the four-component model Fe–S–Si–O probably need not be considered. This suggests that *ab initio* constraints may ultimately turn the problem of core composition from being completely insoluble to being at least partially soluble. □

Received 13 December 1999; accepted 31 March 2000.

- Birch, F. Density and composition of mantle and core. *J. Geophys. Res.* **69**, 4377–4388 (1964).
- Ringwood, A. E. Composition of the core and implications for the origin of the Earth. *Geochim. J.* **11**, 111–135 (1977).
- Poirier, J. P. Light elements in the Earth's outer core: a critical review. *Phys. Earth Planet. Inter.* **85**, 319–337 (1994).
- Loper, D. E. The gravitationally powered dynamo. *Geophys. J. R. Astron. Soc.* **54**, 389–404 (1978).
- Jephcoat, A. & Olson, P. Is the inner core of the Earth pure iron? *Nature* **325**, 332–335 (1987).
- Stixrude, L., Wasserman, E. & Cohen, R. E. Composition and temperature of the Earth's inner core. *J. Geophys. Res.* **102**, 24729–24739 (1997).
- Rama Murthy, V. & Hall, H. T. The chemical composition of the Earth's core: possibility of sulfur in the core. *Phys. Earth Planet. Inter.* **2**, 276–282 (1970).
- Atkins, P. W. *Physical Chemistry* Ch. 7 (Oxford University Press, 1994).
- Boehler, R. Fe–FeS eutectic temperatures to 620 kbar. *Phys. Earth Planet. Inter.* **96**, 181–186 (1996).
- Masters, T. G. & Shearer, P. M. Summary of seismological constraints on the structure of the earth's core. *J. Geophys. Res.* **95**, 21691–21695 (1990).
- Alfè, D. & Gillan, M. J. First-principles simulation of liquid Fe–S under Earth's core conditions. *Phys. Rev. B* **58**, 8248–8256 (1999).
- Alfè, D., Price, G. D. & Gillan, M. J. Oxygen in the Earth's core: A first-principles study. *Phys. Earth Planet. Inter.* **110**, 191–210 (1999).
- Alfè, D., Kresse, G. & Gillan, M. J. Structure and dynamics of liquid iron under Earth's core conditions. *Phys. Rev. B* **61**, 132–142 (2000).
- Ahrens, T. J. Equations of state of iron sulfide and constraints on the sulfur content of the Earth. *J. Geophys. Res.* **84**, 985–998 (1979).

15. de Wijs, G. A., Kresse, G. & Gillan, M. J. First-order phase transitions by first-principles free-energy calculations: The melting of Al. *Phys. Rev. B* **57**, 8223–8334 (1998).
16. Alfe, D., Gillan, M. J. & Price, G. D. The melting curve of iron at the pressures of the Earth's core from *ab initio* calculations. *Nature* **401**, 462–464 (1999).
17. Alfe, D., Gillan, M. J. & Price, G. D. Thermodynamics of hexagonal-close-packed iron under Earth's core conditions. *Phys. Rev. B* (submitted).
18. Alfe, D., de Wijs, G. A., Kresse, G. & Gillan, M. J. Recent developments in *ab-initio* thermodynamics. *Int. J. Quant. Chem.* **77**, 871–879 (2000).
19. Giannozzi, P., de Gironcoli, S., Pavone, P. & Baroni, S. *Ab initio* calculations of phonon dispersions in semiconductors. *Phys. Rev. B* **43**, 7231–7242 (1991).
20. Kresse, G., Furthmüller, J. & Hafner, J. *Ab initio* force-constant approach to phonon dispersion relations of diamond and graphite. *Europhys. Lett.* **32**, 729–734 (1995).
21. Chandler, D. *Introduction to Modern Statistical Mechanics* (Oxford University Press, 1987).
22. Frenkel, D. & Smit, B. *Understanding Molecular Simulation* Ch. 4 (Academic, New York, 1987).
23. Sugino, O. & Car, R. *Ab initio* molecular dynamics study of first-order phase transitions: melting of silicon. *Phys. Rev. Lett.* **74**, 1823–1826 (1995).
24. Kresse, G. & Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
25. Vočadlo, L., Brodholt, J., Alfe, D., Price, G. D. & Gillan, M. J. The structure of iron under the conditions of the Earth's inner core. *Geophys. Res. Lett.* **26**, 1231–1234 (1999).
26. Usselman, T. M. Experimental approach to the state of the core: part I. The liquidus relations of the Fe-rich portion of the Fe-Ni-S system from 30 to 100 kb. *Am. J. Sci.* **275**, 278–290 (1975).
27. Hammond, B. L., Lester, W. A. Jr & Reynolds, P. J. *Monte Carlo Methods in Ab Initio Quantum Chemistry* (World Scientific, Singapore, 1994).
28. Rajagopal, G., Needs, R. J., James, A., Kenny, S. D. & Foulkes, W. M. C. Variational and diffusion quantum Monte Carlo calculations at nonzero wave vectors: Theory and applications to diamond-structure germanium. *Phys. Rev. B* **51**, 10591–10600 (1995).
29. Kent, P. R. C. *et al.* Finite-size errors in quantum many-body simulations of extended systems. *Phys. Rev. B* **59**, 1917–1929 (1999).
30. Kilburn, M. R. & Wood, B. J. Metal-silicate partitioning and the incompatibility of S and Si during core formation. *Earth Planet. Sci. Lett.* **152**, 139–148 (1997).

Acknowledgements

The calculations were performed using the NERC-supported Cray T3E machines at the Manchester CSAR Centre and at the Edinburgh Parallel Computer Centre, and the UCL HiPerSPACE Centre. Support from the NERC and discussions with L. Vočadlo are acknowledged.

Correspondence and requests for materials should be addressed to D.A. (e-mail: d.alfe@ucl.ac.uk).

Detection and classification of atmospheric methane oxidizing bacteria in soil

Ian D. Bull*, Nisha R. Parekh†, Grahame H. Hall‡, Philip Ineson§ & Richard P. Evershed*

* School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK

† Institute of Terrestrial Ecology, Merlewood Research Station, Grange-over-Sands, Cumbria LA11 6JU, UK

‡ Institute of Freshwater Ecology, Far Sawrey, Ambleside, Cumbria LA22 0LP, UK

§ Department of Biology, University of York, P.O. Box 373, York YO10 5YW, UK

Well-drained non-agricultural soils mediate the oxidation of methane directly from the atmosphere, contributing 5 to 10% towards the global methane sink^{1,2}. Studies of methane oxidation kinetics in soil infer the activity of two methanotrophic populations: one that is only active at high methane concentrations (low affinity) and another that tolerates atmospheric levels of methane (high affinity). The activity of the latter has not been demonstrated by cultured laboratory strains of methanotrophs, leaving the microbiology of methane oxidation at atmospheric concentrations unclear^{3,4}. Here we describe a new pulse-chase experiment using long-term enrichment with ¹²CH₄ followed by short-term exposure to ¹³CH₄ to isotopically label methanotrophs in a soil from a temperate forest. Analysis of labelled phospholipid fatty acids (PLFAs) provided unambiguous evidence of methane assimilation at true atmospheric concentrations (1.8–3.6 p.p.m.v.). High proportions of ¹³C-labelled C₁₈ fatty acids and

the co-occurrence of a labelled, branched C₁₇ fatty acid indicated that a new methanotroph, similar at the PLFA level to known type II methanotrophs, was the predominant soil micro-organism responsible for atmospheric methane oxidation.

Culturable methane oxidizing bacteria are classified into types I, II and X depending on the guanine and cytosine content of their DNA, intracellular membrane arrangement, carbon assimilation pathway and PLFA composition³. Active soil microbial populations utilizing a ¹³C-labelled substrate will readily incorporate ¹³C into membrane lipid components such as PLFAs. Recent work has demonstrated the power of ¹³C-labelling to link lipid biomarkers to other biogeochemical processes in lake and marine sediments^{5,6}. We have applied similar techniques to a soil to investigate the oxidation of methane at atmospheric concentrations.

Soils used for pulse-chase labelling were obtained from a Sitka Spruce plantation in North Wales, UK. Mull-board ploughing before tree planting in 1959 produced a repeating microtopographic pattern consisting of a plough furrow and ridge followed by an undisturbed ridge. The plough ridges have an inversion of the original stratification of soil horizons resulting in a buried organic horizon beneath an overturned mineral horizon; soil cores from these ridges exhibited high rates of atmospheric methane consumption (mean rate is 77 µg m⁻² h⁻¹, standard error (s.e.) = 2, n = 7). Laboratory experiments at atmospheric and elevated concentrations demonstrated that soils from both the buried organic and mineral horizons showed similar kinetics of oxidation, but activities were generally lower in the mineral soil. Increasing the initial concentration of methane from 1.8 to 300 parts per million by volume (p.p.m.v.) for both soils showed typical Michaelis–Menton kinetics with K_m methane concentration at half of the maximal reaction rate values of 25 and 56 p.p.m.v. for the buried organic and mineral soils, respectively (Fig. 1). Rapid rates of methane consumption, characteristic of high-affinity methane-oxidation kinetics, were observed in both soils below 50 p.p.m.v. methane. Concentrations of methane, which induce oxidation kinetics indicative of high (1.8–3.6 p.p.m.v.) and low (100 p.p.m.v.) affinity methanotroph populations, were chosen for the ¹³C pulse-chase experiment in soil columns.

The distribution of PLFAs shown in Fig. 2 (for the buried organic horizon) is typical of all the soil samples analysed; the only significant difference between the two soil types is the lower overall abundance of PLFAs in the mineral soil (~35 µg g⁻¹ TOC); where TOC refers to 'total organic carbon' compared with the buried organic

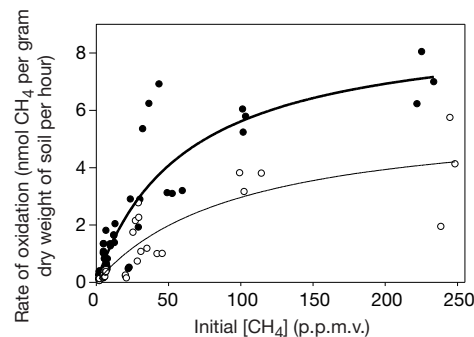


Figure 1 Rates of methane oxidation between 1.5 and 250 p.p.m.v. by forest soil from the mineral and buried organic layers. Lines represent direct nonlinear regression analysis and show typical Michaelis–Menton kinetics. Data points show the mineral layer (white circles) and buried organic layer (black circles). Application of the Lineweaver Burke plot provided a significant regression ($P < 0.001$) for both data sets that were used to calculate the apparent K_m (methane concentration at half of the maximal reaction rate) and V_{max} (maximal reaction rate) values for each soil. Regression for mineral soil $R^2 = 0.78$, apparent $K_m = 56$ p.p.m.v., apparent $V_{max} = 2.8$ nmol per gram dry weight of soil per hour. Regression for organic soil $R^2 = 0.82$, apparent $K_m = 25$ p.p.m.v., apparent $V_{max} = 3.7$ nmol per gram dry weight of soil per hour.

soil ($\sim 75 \mu\text{g g}^{-1}_{\text{TOC}}$). This quantitative difference has been reported previously and arises from the inherently higher biomass sustained by the nutrient-rich buried organic soil⁷. This is also reflected in the higher rates of methane oxidation, and therefore presumably greater biomass, in the buried organic horizon. PLFAs in both the buried organic and mineral soils comprise mainly C_{16} and C_{18} fatty acids and apart from the saturated components are composed predominantly of 1 ω 7 (indicative of Gram negative aerobes, obligate anaerobes) and 1 ω 9 (indicative of Gram positive bacteria) unsaturated compounds⁸. Significant quantities of 18:1 ω 8c unsaturates, considered highly indicative of type II methanotrophs such as *Methylosinus trichosporium*, are not detectable in any of the soils (Fig. 2). Although used routinely for studies of soil biomass and bacterial community structure, the drawbacks of quantitative lipid profiling are quickly realized when attempting to assess the microbial populations involved in specific processes within the same soil. No significant differences can be detected between the PLFA compositions of soils exposed to 1.8–3.6 p.p.m.v. methane and those incubated under elevated methane (100 p.p.m.v.; results not shown) because methanotrophs are minor contributors to the overall soil microbial biomass. However, incorporation and subsequent detection of a stable isotope label by a pulse-chasing approach using a process substrate, for example methane ($^{13}\text{CH}_4$), easily overcomes limitations associated with the profiling method. In this way gas chromatography combustion isotope ratio mass spectrometry provides a uniquely specific and sensitive means of determining any metabolic incorporation of the ^{13}C label into PLFAs derived from populations of methane-oxidizing bacteria.

Analysis of PLFAs from the buried organic soil exposed to atmospheric levels of methane (1.8–3.6 p.p.m.v.) clearly shows incorporation of the ^{13}C label in the br17:0 (component 16) and 18:1 ω 7c (component 27) PLFAs (Fig. 3a). This indicates that the oxidation of atmospheric methane in this soil is mediated primarily by organisms with a labelling pattern similar to that of a type II methanotroph^{8,9}. The presence of type II *Methylosinus/Methylocystis*

strains, although not the species *Methylosinus trichosporium*, could possibly account for the heavily labelled 18:1 ω 7c component, the lack of labelled 18:1 ω 8c and the low number of other ^{13}C labelled PLFAs^{8,9}. Indeed, this would be entirely consistent with previous work which has utilized gene cloning and sequencing techniques to identify the presence of this phylogenetic group in a four-year-long liquid enrichment culture under elevated concentrations of methane (<275 p.p.m.v.; ref. 10). However, the co-occurrence of an equally labelled br17:0 PLFA in the soil samples is not consistent with the PLFA profiles of previously described type II methanotrophs. Trace amounts of a br17:0 component have been detected in *Methylobacterium organophilum* and *Methylosinus trichosporium* as well as in natural-gas-enriched soils^{7–9,11,12}. The data shows that although the br17:0 component constitutes a small proportion of the total PLFAs in all of the methane-treated soils (6.4–7.8%), the degree of incorporation of carbon from $^{13}\text{CH}_4$ in this component is extremely high compared with that in the 16:1 and 18:1 signature lipids for known methanotrophic bacteria. This supports the proposition that methane oxidation is mediated predominantly by a new methanotroph which expresses some similarity at the PLFA level to known type II methanotrophs.

The buried organic soil exposed to elevated levels of methane (100 p.p.m.v.) exhibits significant ^{13}C -labelling of the 16:0, 16:1 ω 8c,

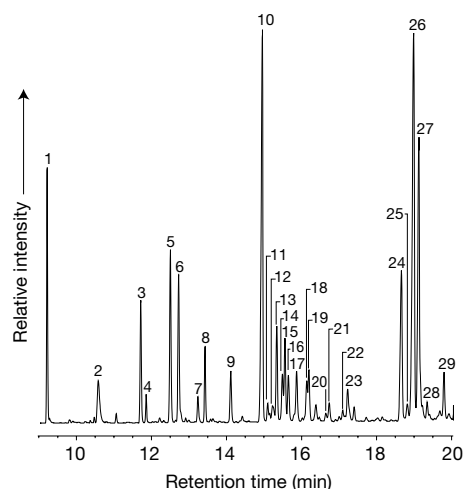


Figure 2 A partial gas chromatogram of the methylated phospholipid fatty acid (PLFA) fraction derived from the buried organic soil exposed to an atmospheric (1.8–3.6 p.p.m.v.) concentration of methane. Peak numbers indicate PLFAs as follows: peak number 1 is 12:0; 2 is nonadecane (internal standard; IS); 3 is 14:0; 4 is 8:0 diacid; 5 is 15:0; 6 is 15:0; 7 is 15:0; 8 is 9:0 diacid; 9 is br16:0; 10 is 16:0; 11 is 16:1 ω 11; 12 is 16:1 ω 8c; 13 is 16:1 ω 7c; 14 is br17:0; 15 is 16:1 ω 5; 16 is br17:0; 17 is 17:0; 18 is 17:0; 19 is br17:1; 20 is br18:0; 21 is 17:0; 22 is 17:1 ω 8; 23 is br18:0; 24 is 18:0; 25 is 18:1 ω 13; 26 is 18:1 ω 9c; 27 is 18:1 ω 7c; 28 is 18:1 ω 5c; 29 is 18:2; where br/a/i indicates a branched/iso/anteiso alkyl chain; the number preceding a colon indicates the total carbon number; the number after the colon indicates the number of double bonds; the number following ω indicates the position of unsaturation counting from the terminal methyl group; and c or t indicates a *cis* or *trans* geometric isomer.

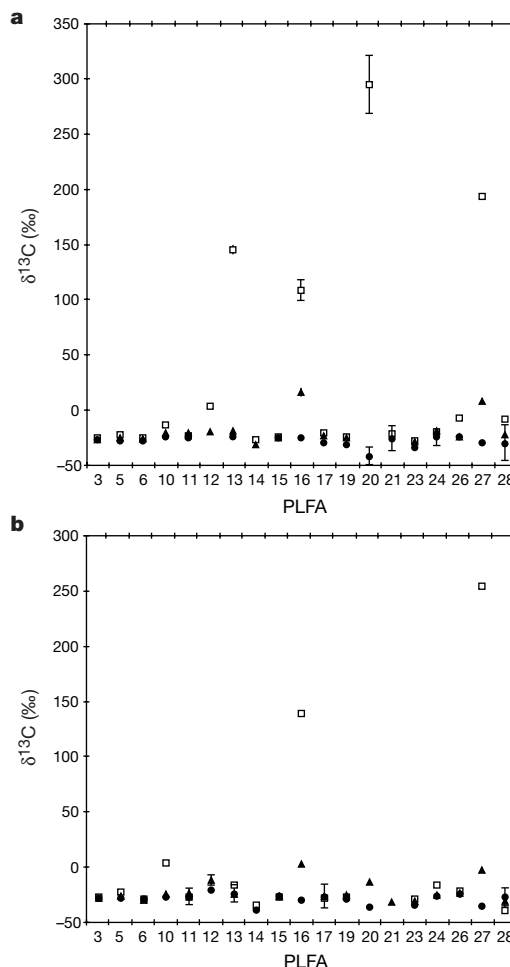


Figure 3 A plot of $\delta^{13}\text{C}$ values of PLFAs from soils after incubation and $^{13}\text{CH}_4$ pulse-labelling. **a**, Three replicate buried organic soils. **b**, Three replicate mineral soils. Pulse-labelling was done under elevated methane (100 p.p.m.v., white squares), atmospheric methane (1.8–3.6 p.p.m.v., black triangles) and control (non-enriched and unlabelled soil, black circles). Missing data points correspond to peaks for which a reliable measurement could not be obtained. Numbers on the x axes correspond to the PLFAs listed in the caption to Fig. 2.

16:1 ω 7c, br17:0, br18:0, 18:1 ω 9c and 18:1 ω 7c components (Fig. 3a). The range is limited and indicates an absence of any significant recycling and uptake of ^{13}C through $^{13}\text{CO}_2$ by autotrophic bacteria during the course of the experiment. Such high levels of incorporation in both C_{16} and C_{18} PLFAs suggest that low-affinity methane oxidation in this soil is being performed by organisms related to both type I and type II methanotrophs¹¹. In addition, the small number of heavily labelled components indicates that this process is still, even under high concentrations of methane, being mediated by a limited diversity of bacterial species. The labelled type I signature (16:1) PLFAs exhibited by the soil are closely matched by those present in *Methylomonas gracilis*, indicating that either this, or a closely related species, may be partially responsible for low-affinity methanotrophy in this soil¹². When soil is exposed to elevated concentrations of methane, the activity of low-affinity species will increase and the activity of high-affinity species will continue. Therefore, the same type II species operating at atmospheric concentrations of methane are likely to constitute a significant proportion of the methanotrophic population in the soils exposed to elevated methane¹³.

Stable isotope analyses of individual PLFAs from mineral soil exposed to both atmospheric (1.8–3.6 p.p.m.v.) and elevated (100 p.p.m.v.) levels of methane show incorporation of the ^{13}C -label in the 16:0, br17:0 and 18:1 ω 7c PLFAs (Fig. 3b). Comparison with the profiles of reference bacteria suggests a population of methanotrophs dominated by an unknown organism, which is possibly related to the type II *Methylosinus*/*Methylocystis* genera⁸. This contrasts with the active methanotrophic population in the buried organic soil exposed to elevated methane, which included organisms related to both type I and type II methanotrophs. A similar population of high-affinity methane-oxidizing bacteria, dominated by new organisms related to type II genera, are present in both the buried organic and mineral horizons of this forest soil.

The ability to characterize active populations of bacteria *in situ* at true atmospheric concentrations of methane facilitates a reliable and ecologically relevant understanding of soil methane oxidation. Alternative DNA sequencing techniques lack sensitivity and require artificially high concentrations of methane¹⁰, whereas radiocarbon labelling with ^{14}C is also compromised by a lack of both sensitivity and compound specificity¹⁴. Compound-specific stable-isotope analysis provides a sensitive and non-hazardous means of investigating the activity and roles of microorganisms involved in important biogeochemical cycles. In addition, biologically relevant signature compounds, other than PLFAs, may be studied (for example, non-standard amino acids, amino sugars, hopanoids, ergosteroids and mycolipids). Information about soil methanotrophs can be used to formulate land management strategies with the long-term goal of controlling and maximizing the quantity of methane oxidized by terrestrial biomes. □

Methods

Methane-oxidation potentials

Soil cores, of internal diameters 15 and 30 cm, were collected from plough ridges on the floor of a temperate Sitka Spruce forest called Coed Anafon situated on the North Wales Coast (53°12' N, 4°00' W). The soil cores were separated into three horizons; the upper organic, the mineral, and the buried organic horizons. Samples of both the buried organic and mineral soil horizons were sieved (2-mm mesh) and stored at 4 °C until required (<6 months). For the kinetic studies, the soil samples were acclimatized overnight at 20 °C, thoroughly mixed and triplicate 10-g samples placed in serum bottles to which 10 ml of sterile water was added. The bottles were sealed using butyl rubber septa and aluminium crimps. Incubation was at 20 °C on a shaking table set at 150 r.p.m. The rate of methane oxidation was estimated from the decrease in headspace concentration over 60 to 120 min to ensure that instantaneous rates were determined. The short incubation measures the activity of methane monooxygenase enzyme present in the sample and not the activity that develops after exposure to added methane. At least three samples of headspace gas were removed during the incubation period. The methane concentration in the headspace was changed by adding standard methane/air gas mixtures. The concentration range for each soil horizon was varied from 1.7 to 8,000 p.p.m.v. methane. Methane concentrations were determined by gas chromatography¹⁵.

^{13}C -methane labelling

Glass sinter tubes with 20 g dry weight equivalent of soil from the buried organic or mineral horizons were used as microcosms and incubated in a 20 °C constant temperature room. A peristaltic pump was used to flow methane in air, at atmospheric (1.8 p.p.m.v.) or elevated (100 p.p.m.v.) concentrations, and sterile distilled water through the microcosms at constant rates of 0.05 ml $\text{CH}_4 \text{ min}^{-1}$ and 0.015 ml $\text{H}_2\text{O min}^{-1}$, respectively, over a six-month incubation period. The microcosms were sealed only at the top to allow for drainage and venting of excess gas. The headspace gases over the equilibrated soils in the atmospheric and elevated methane microcosms were then respectively replaced with 1.8 or 100 $^{13}\text{CH}_4$ in air. Microcosms were incubated for a further three weeks to label the active methanotrophic bacteria populations within these soils. Thus, the microcosms used to simulate atmospheric methane oxidation contained no more than 3.6 p.p.m.v. methane at any stage of the experiment. $^{13}\text{CH}_4$ was obtained from Isotech Inc., USA and graded 99.9% pure and 99.4% ^{13}C atom enriched. At the end of the equilibration and ^{13}C -labelling period, sub-samples of each soil were freeze-dried and PLFAs were extracted from these samples.

PLFA analysis

Lipids were extracted by following a modified Bligh and Dyer method^{16,17}. Subsequent fractionation on a silica column with a bonded amino-propyl phase yielded a total PLFA fraction. Fatty acid components were released by mild alkaline hydrolysis (0.1 M KOH in MeOH) and then isolated and methylated using a 14% (w/v) solution of a boron trifluoride-methanol complex. Individual compounds were quantified using an HP 5890 series II gas chromatograph equipped with a Chrompack CPWax-52CB (50 m length, 0.32 mm internal diameter, 0.12 μm film thickness; H_2 carrier gas, 10 psi head-pressure). Compound structures were assigned unambiguously by gas chromatography/mass spectrometry through the analysis of dimethyldisulphide adducts of the individual mono-unsaturated PLFAs. Identifications were made using a Carlo Erba 5160 gas chromatograph (column as above; He carrier gas, 10 psi head-pressure) coupled, using a heated transfer line (320 °C), to a Finnigan MAT 4500 quadrupole mass spectrometer scanning in the range of m/z 50 to 650 with a cycle time of 1.0 s (current was maintained at 300 μA with an ion source temperature of 190 °C and an electron voltage of 70 eV). Stable carbon-isotope compositions of individual compounds were determined using a Finnigan Delta S isotope mass spectrometer coupled to a Varian 3400 gas chromatograph (column as above; He carrier gas, 15 psi head-pressure) with an extensively modified Finnigan type I combustion interface. $\delta^{13}\text{C}$ values and associated errors were determined after correcting for the isotopic composition of the exogenous methyl derivative group^{18,19}. Column temperatures were programmed from 40 °C (1 min isothermal) to 150 °C at a rate of 15 °C min^{-1} , and then to 240 °C at a rate of 4 °C min^{-1} with a final isothermal period of 20 min.

Received 4 January 1999; accepted 18 February 2000.

1. Cicerone, R. J. & Oremland, R. S. Biogeochemical aspects of atmospheric methane. *Glob. Biogeochem. Cycles* **2**, 299–327 (1988).
2. Crutzen, P. J. Methane's sinks and sources. *Nature* **350**, 380–381 (1991).
3. Hanson, R. S. & Hanson, T. E. Methanotrophic bacteria. *Microbiol. Rev.* **60**, 439–471 (1996).
4. Conrad, R. Soil microorganisms as controllers of atmospheric trace gases (H_2 , CO , CH_4 , OCS , N_2O , and NO). *Microbiol. Rev.* **60**, 609–643 (1996).
5. Boschker, H. T. S. *et al.* Direct linking of microbial populations to specific biogeochemical processes by ^{13}C -labelling of biomarkers. *Nature* **392**, 801–804 (1998).
6. Hinrichs, K.-U., Hayes, J. M., Sylva, S. P., Brewer, P. G. & Delong, E. F. Methane-consuming archaeobacteria in marine sediments. *Nature* **398**, 802–805 (1999).
7. Zelles, L. & Bai, Q. Y., Rackwitz, R., Chadwick, & Beese, F. Determination of phospholipid and lipopolysaccharide-derived fatty acids as an estimate of microbial biomass and community structures in soils. *Biol. Fertil. Soils* **19**, 115–123 (1995).
8. Bowman, J. P., Skerratt, J. H., Nichols, P. D. & Sly, L. I. Phospholipid fatty acid and liposaccharide fatty acid signature lipids in methane utilizing bacteria. *FEMS Microbiol. Ecol.* **85**, 15–22 (1991).
9. Nichols, P. D., Smith, G. A., Antworth, C. P., Hanson, R. S. & White, D. C. Phospholipid and lipopolysaccharide normal and hydroxy fatty acids as potential signatures for methane-oxidizing bacteria. *FEMS Microbiol. Ecol.* **31**, 327–335 (1985).
10. Dunfield, P. F., Werner, L., Henckel, T., Knowles, R. & Conrad, R. High affinity methane oxidation by a soil enrichment culture containing a type II methanotroph. *Appl. Environ. Microbiol.* **65**, 1009–1014 (1999).
11. Nichols, P. D. *et al.* Detection of a microbial consortium, including type II methanotrophs, by use of phospholipid fatty acids in an aerobic halogenated hydrocarbon-degrading soil column enriched with natural gas. *Environ. Toxicol. Chem.* **6**, 89–97 (1987).
12. Guckert, J. B., Ringelberg, D. B., White, D. C., Hanson, R. S. & Bratina, B. J. Membrane fatty acids as phenotypic markers in the polyphasic taxonomy of methylotrophs within the Proteobacteria. *J. Gen. Microbiol.* **137**, 2631–2641 (1991).
13. Bender, M. & Conrad, R. Effect of CH_4 concentrations and soil conditions on the induction of CH_4 oxidation activity. *Soil Biol. Biochem.* **27**, 1517–1527 (1995).
14. Roslev, P. & Iversen, N. Radioactive fingerprinting of microorganisms that oxidize atmospheric methane in different soils. *Appl. Environ. Microbiol.* **65**, 4064–4070 (1999).
15. Hall, G. H., Simon, B. M. & Pickup, R. W. CH_4 production in blanket bog peat: a procedure for sampling, sectioning and incubating samples whilst maintaining anaerobic conditions. *Soil Biol. Biochem.* **28**, 9–15 (1996).
16. Bligh, E. G. & Dyer, W. J. A rapid method of total lipid extraction and purification. *Can. J. Biochem. Physiol.* **37**, 911–917 (1959).
17. Guckert, J. B., Antworth, C. P., Nichols, P. D. & White, D. C. Phospholipid, ester-linked fatty-acid profiles as reproducible assays for changes in prokaryotic community structure of estuarine sediments. *FEMS Microbiol. Ecol.* **31**, 147–158 (1985).
18. Jones, D. M., Carter, J. F., Eglinton, G., Jumeau, E. J. & Fenwick, C. S. Determination of $\delta^{13}\text{C}$ values of straight chain and cyclic alcohols by gas chromatography isotope ratio mass spectrometry. *Biol. Mass Spectrom.* **20**, 641–646 (1991).

19. Rieley, G. Derivatization of organic compounds prior to gas chromatographic-combustion-isotope ratio mass spectrometric analysis: Identification of isotope fractionation processes. *Analyst* **119**, 915–919 (1994).

Acknowledgements

We thank J. Carter and A. Gledhill for their help with GC/MS and GCC/IRMS analyses and C. Davies for setting up the long-term enrichment experiment. This study was supported by the Natural Environment Research Council through the CEH Integrated Fund (N.R.P., P.I., G.H.H.) and through a Non-thematic Research Grant (P.I. and R.P.E.).

Correspondence and requests for materials should be addressed to R. P. E. (e-mail: r.p.evershed@bris.ac.uk).

Infectious parthenogenesis

M. E. Huigens*, R. F. Luck†, R. H. G. Klaassen*, M. F. P. M. Maas*, M. J. T. N. Timmermans* & R. Stouthamer*

* Laboratory of Entomology, Department of Plant Sciences Wageningen University, 6700 EH Wageningen, The Netherlands

† Department of Entomology, University of California, Riverside, California 92521, USA

Parthenogenesis-inducing *Wolbachia* bacteria are reproductive parasites that cause infected female wasps to produce daughters without mating^{1,2}. This manipulation of the host's reproduction enhances the transmission of *Wolbachia* to future generations because the bacteria are passed on vertically only from mothers to daughters. Males are dead ends for cytoplasmically inherited bacteria: they do not pass them on to their offspring. Vertical transmission of *Wolbachia* has been previously considered to be the main mode of transmission. Here we report frequent horizontal transmission from infected to uninfected wasp larvae sharing a common food source. The transferred *Wolbachia* are then vertically transmitted to the new host's offspring. This natural and unexpectedly frequent horizontal transfer of parthenogenesis-inducing *Wolbachia* intraspecifically has important implications for the co-evolution of *Wolbachia* and their host.

Trichogramma kaykai (Hymenoptera; Trichogrammatidae) is a minute egg parasitoid of a butterfly species, *Apodemia mormo deserti* (Lepidoptera; Lycaenidae), which inhabits the Mojave Desert of southern California³. *T. kaykai* normally lays a clutch of 3–5 offspring in an *A. mormo deserti* egg. This parasitoid wasp manifests haplo-diploid sex determination in which unfertilized eggs develop into sons and fertilized eggs develop into daughters. Between 6% and 26% of the *T. kaykai* females in these populations are infected with parthenogenesis-inducing *Wolbachia* and produce daughters from their unfertilized eggs⁴. These daughters are produced when parthenogenesis-inducing *Wolbachia* suppresses spindle formation during anaphase of the first mitotic division, thus restoring diploidy by fusion of the two mitotic nuclei⁵. In the field, the offspring of two *T. kaykai* females can share and mature in the same butterfly egg⁶. Thus, we tested whether *Wolbachia* transfers horizontally between the offspring of a *Wolbachia*-infected and an uninfected *T. kaykai* female when they share the same egg and, if so, whether this infection persists in its expression in subsequent generations.

T. kaykai females collected at Last Chance Canyon, El Paso Mountains, Kern County, California, were used to initiate infected and uninfected lines. We cultured these lines in the laboratory on eggs of the moth *Trichoplusia ni*⁷, which do not harbour *Wolbachia* (R. S., unpublished data). Each of these lines differed in the size of a microsatellite DNA repeat. In our experiments, we offered a moth egg to a female *T. kaykai* from one line, allowing her to lay a normal clutch of 2–4 eggs. Two hours later, we exposed the same egg to a

female from the second line. In half the cases, a *Wolbachia*-infected *T. kaykai* was offered the egg first and in the other half the order was reversed. We continuously observed and recorded the number of eggs oviposited in a moth egg by each *T. kaykai*⁸. The resulting offspring were linked unambiguously to their parental female using a microsatellite marker. If F₁ offspring from a doubly parasitized egg consisted of only females, we exposed these virgin F₁ females individually to host eggs and recorded the sex of their progeny. Because only infected virgins produce daughters, their presence in offspring of an F₁ virgin originating from an uninfected line indicated horizontal transfer of parthenogenesis-inducing *Wolbachia*.

We found clear evidence for the horizontal transmission of parthenogenesis-inducing *Wolbachia* and the expression of parthenogenesis. Twenty-one instances of horizontal transfer occurred in 56 all-female broods from moth eggs that contained infected and uninfected larvae (Table 1). Horizontal transfer to uninfected offspring occurred independent of the order in which the infected and uninfected females were exposed to the host egg ($\chi^2_{0.05,1} = 2.75$; $n = 56$). The bacteria's presence was confirmed in all newly infected females by means of polymerase chain reaction (PCR) using *Wolbachia*-specific primers⁹. Both sons and daughters were produced by these newly infected, virgin F₁ females, indicating inefficient vertical transmission from mother to offspring. We suspect that these females had insufficient *Wolbachia* titre to express parthenogenesis in all of their offspring. To test whether a new infection was maintained over several generations, we established three lines, each of which consisted of offspring from a different, newly infected female. Vertical transmission was inefficient in the first generation, but it was highly effective in the subsequent eight generations. The transmission efficiency was 100% for 10 virgin females of each line tested in generation 3 and generation 8.

The process by which uninfected *Trichogramma* larvae acquire *Wolbachia* remains unclear. Several possible acquisition mechanisms exist. Infected larvae may die and be consumed by uninfected larvae. Larvae may fight within the host egg and the uninfected acquire an infection through blood-to-blood contact as found in the woodlice *Armadillidium vulgare*¹⁰. Finally, an infected female may inject *Wolbachia* into the host egg when she oviposits in it and the larvae acquire the infection either orally or through wounds caused by larval conflict. The first possibility can be eliminated: in 3 out of the 21 cases, all the eggs laid by the infected female survived to become adults, yet some of the offspring from the uninfected female became infected. We cannot distinguish between the two other possibilities.

Our results show that offspring of uninfected females can acquire sufficient parthenogenesis-inducing *Wolbachia* to express parthenogenesis when they share a butterfly egg with offspring of an infected female. We know that such egg sharing occurs in the field. The maximum rates at which the offspring of *Wolbachia*-infected and uninfected females share the same butterfly egg in the field can be estimated from observed parasitism rates¹¹. *T. kaykai* can parasitize up to 80% of the *Apodemia mormo deserti* eggs, suggesting that offspring of two or more females could share the same egg in as many as 48% of cases. Up to 9% of the latter cases would involve offspring of both infected and uninfected females (assuming a 10% infection rate among females). Although such high parasitism rates are infrequent, they illustrate the potential for horizontal transfer under field conditions.

Purely vertically inherited symbionts suffer from an accumulation of mutations similar to the Muller's ratchet in asexual organisms¹². However, horizontal transfer of parthenogenesis-inducing *Wolbachia* among offspring of infected females offers the potential for recombination between different *Wolbachia* variants and thus for circumventing Muller's ratchet.

Vertical transmission has been considered the only common mechanism by which parthenogenesis-inducing *Wolbachia* is trans-

Table 1 Twenty-one cases of horizontal transfer of parthenogenesis-inducing *Wolbachia* between *T. kaykai* wasps parasitizing the same host egg

Host egg	P		F ₁	F ₂		♀3	♀4
	No of eggs oviposited (size of microsatellite fragment)			♀1	♀2		
1	3 (197) U	2 (203) I	→	♀ ♀ ♀	1 ♀:2 ♂ (197) U	1 ♀:0 ♂ (203) I	
2	2 (216) I	2 194 U	→	♀ ♀ ♀ ♀	17 ♀:0 ♂ (216) I	15 ♀:0 ♂ (216) I	4 ♀:10 ♂ (194) U
3	2 (194) U	1 (216) I	→	♀ ♀	18 ♀:0 ♂ (216) I	9 ♀:14 ♂ (194) U	
4	2 (222) I	2 (197) U	→	♀ ♀	9 ♀:5 ♂ (197) U	10 ♀:0 ♂ (222) I	
5	2 (222) I	2 (197) U	→	♀ ♀	1 ♀:8 ♂ (197) U	9 ♀:8 ♂ (197) U	
6	2 (194) U	1 (222) I	→	♀ ♀	9 ♀:7 ♂ (194) U	14 ♀:19 ♂ (194) U	
7	2 (200) U	1 (216) I	→	♀ ♀	0 ♀:14 ♂ (200) U	4 ♀:3 ♂ (200) U	
8	2 (200) U	1 (216) I	→	♀ ♀	0 ♀:12 ♂ (200) U	1 ♀:14 ♂ (200) U	
9	2 (200) U	2 (216) I	→	♀ ♀	1 ♀:11 ♂ (200) U	0 ♀:20 ♂ (200) U	
10	2 (197) U	2 (203) I	→	♀ ♀	3 ♀:14 ♂ (197) U	3 ♀:0 ♂ (203) I	
11	2 (197) U	2 (203) I	→	♀ ♀	7 ♀:2 ♂ (197) U	1 ♀:6 ♂ (197) U	
12	2 (197) U	2 (203) I	→	♀ ♀	4 ♀:5 ♂ (197) U	3 ♀:6 ♂ (197) U	
13	2 (216) I	1 (194) U	→	♀ ♀	6 ♀:2 ♂ (194) U	7 ♀:0 ♂ (216) I	
14	2 (216) I	1 (194) U	→	♀ ♀	5 ♀:21 ♂ (194) U	DNE	
15	2 (194) U	2 (216) I	→	♀ ♀	11 ♀:2 ♂ (194) U	1 ♀:18 ♂ (194) U	
16	2 (194) U	2 (216) I	→	♀ ♀	3 ♀:25 ♂ (194) U	5 ♀:14 ♂ (194) U	
17	2 (194) U	1 (216) I	→	♀ ♀	6 ♀:1 ♂ (194) U	DNE	
18	2 (194) U	1 (219) I	→	♀ ♀	1 ♀:37 ♂ (194) U	DNE	
19	2 (194) U	1 (219) I	→	♀ ♀	9 ♀:1 ♂ (194) U	1 ♀:29 ♂ (194) U	
20	2 (194) U	2 (219) I	→	♀ ♀	4 ♀:3 ♂ (194) U	3 ♀:5 ♂ (194) U	
21	2 (219) I	1 (194) U	→	♀ ♀	5 ♀:5 ♂ (194) U	DNE	

Bold entries indicate *Wolbachia* infection. The origin of the wasps is mentioned: U, uninfected; I, infected. DNE, did not emerge.

mitted. Such transmission should select for accommodation between *Wolbachia* and its host wasps; however, when horizontal transfer and vertical transmission are frequent, they should select for conflicting adaptations because they are considered to require opposing adaptations¹³. The reduced offspring production by parthenogenesis-inducing *Wolbachia* females¹⁴ may be a result of such opposing adaptations.

Intraspecific horizontal transfer of *Wolbachia* from infected to uninfected larvae within a host suggests the possibility of intra-specific transfers. These transfers have been predicted on the basis of phylogenies of *Wolbachia* and its Trichogrammatid hosts¹⁵. Host eggs that are shared by offspring of more than one *Trichogramma* species have been found in field populations in the Mojave desert (M.E.H. and R.S., unpublished data). One of these species, *Trichogramma deion*, is known to be infected, but with a phylogenetically different *Wolbachia* isolate¹⁶. Thus, horizontal transfer between species may be more difficult or infection may not be maintained in the new species. Such interspecific horizontal transfers, however, may offer a means with which to investigate the co-evolution between *Wolbachia* and its host wasp. □

Received 9 November 1999; accepted 17 February 2000.

1. Stouthamer, R., Luck, R. F. & Hamilton, W. D. Antibiotics cause parthenogenetic *Trichogramma* to revert to sex. *Proc. Natl Acad. Sci. USA* **87**, 2424–2427 (1990).
2. Stouthamer, R., Breuwer, J. A. J., Luck, R. F. & Werren, J. H. Molecular identification of microorganisms associated with thelytoky. *Nature* **361**, 66–68 (1993).
3. Pinto, J. D., Stouthamer, R. & Platner, G. R. A new cryptic species of *Trichogramma* from the Mojave desert of California as determined by morphological, reproductive and molecular data. *Proc. Entomol. Soc. Washington* **99**, 238–247 (1997).

4. Stouthamer, R. & Kazmer, D. J. Cytogenetics of microbe-associated parthenogenesis and its consequences for gene flow in *Trichogramma* wasps. *Heredity* **73**, 317–327 (1994).
5. Stouthamer, R. in *Influential Passengers: Inherited Microorganisms and Arthropod Reproduction* (eds O'Neill, S. L., Hoffmann, A. A. & Werren, J. H.) 102–122 (Oxford Univ. Press, Oxford, 1997).
6. Kazmer, D. J. *The Evolution of Reproductive Strategies in Trichogramma and its Relations to Biological Control*. Thesis, Univ. California, Riverside (1992).
7. Pak, G. A. & Oatman, E. R. Comparative life table, behavior and competition studies of *Trichogramma brevicapillum* and *T. pretiosum*. *Entomol. Exp. Appl.* **32**, 68–79 (1982).
8. Suzuki, Y., Tsuji, H. & Sasakawa, M. Sex allocation and the effects of superparasitism on secondary sex ratios in the gregarious parasitoid, *Trichogramma chilonis*. *Anim. Behav.* **32**, 478–484 (1984).
9. Braig, H. R., Zhou, W., Dobson, S. & O'Neill, S. L. Cloning and characterization of a gene encoding the major surface protein of the bacterial endosymbiont *Wolbachia*. *J. Bacteriol.* **180**, 2373–2378 (1998).
10. Rigaud, T. & Juchault, P. Success and failure of horizontal transfers of feminizing *Wolbachia* endosymbionts in woodlice. *J. Evol. Biol.* **8**, 249–255 (1995).
11. Alphen, J. J. M. van, Visser, M. E. & Nell, H. W. Adaptive superparasitism and patch time allocation in solitary parasitoids: searching in groups vs sequential patch visits. *Funct. Ecol.* **6**, 528–535 (1992).
12. Moran, N. A. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA* **93**, 2873–2878 (1996).
13. Ewald, P. W. *Evolution of Infectious Disease* (Oxford Univ. Press, Oxford, 1994).
14. Stouthamer, R. & Luck, R. F. Influence of microbe-associated parthenogenesis on the fecundity of *Trichogramma deion* and *Trichogramma pretiosum*. *Entomol. Exp. Appl.* **67**, 183–192 (1993).
15. Schilthuisen, M. & Stouthamer, R. Horizontal transmission of parthenogenesis-inducing microbes in *Trichogramma* wasps. *Proc. R. Soc. Lond. B* **264**, 361–366 (1997).
16. Schilthuisen, M., Honda, J. & Stouthamer, R. Parthenogenesis-inducing *Wolbachia* in *Trichogramma kaykai* originates from a single infection. *Ann. Entomol. Soc. America* **91**, 410–414 (1998).

Acknowledgements

This work was supported by Earth and Life Sciences Foundation (ALW) of the Netherlands (M.E.H. and R. S.) and by a NATO collaborative grant (R.S. and R.F.L.). We thank B. Koopmanschap for her assistance with the molecular work.

Correspondence and requests for materials should be addressed to R.S. (e-mail: Richard.Stouthamer@users.ento.wau.nl).

Rewarding effects of opiates are absent in mice lacking the receptor for substance P

Patricia Murtra^{*}, Anne M. Sheasby[†], Stephen P. Hunt[†] & Carmen De Felipe^{*}

^{*} Instituto de Neurociencias, Universidad Miguel Hernandez, Ap. correos 18, 03550 Alicante, Spain

[†] Department of Anatomy and Developmental Biology, University College London, Medawar Building, Gower Street, London WC1E 6BT, UK

Modulation of substance P activity offers a radical new approach to the management of depression, anxiety and stress^{1–3}. The substance P receptor is highly expressed in areas of the brain that are implicated in these behaviours, but also in other areas such as the nucleus accumbens which mediate the motivational properties of both natural rewards such as food and of drugs of abuse such as opiates^{4–7}. Here we show a loss of the rewarding properties of morphine in mice with a genetic disruption of the substance P receptor. The loss was specific to morphine, as both groups of mice responded when cocaine or food were used as rewards. The physical response to opiate withdrawal was also reduced in substance P receptor knockout mice. We conclude that substance P has an important and specific role in mediating the motivational aspects of opiates and may represent a new pharmacological route for the control of drug abuse.

The rewarding and addictive properties of opiates are thought to depend upon both dopaminergic and non-dopaminergic mechanisms within the ventral forebrain^{6–12}. We investigated whether disruption of the substance P receptor gene in mice would affect motivational systems and addictive processes by examining the effects of morphine or cocaine on locomotion and conditioned place preference. Initially, motor behaviour was assessed in a variety of different environments and substance P receptor knockout (*NK1*^{−/−}) mice were found to be indistinguishable from wild-type mice¹. Behaviour in an elevated plus maze and exploratory behaviour were unchanged (Fig. 1a, b). All mice habituated to a new environment upon repeated exposure (Fig. 1b). Thus, locomotor activity in the *NK1*^{−/−} mouse and the response to aversive environmental conditions were comparable to those of wild-type mice.

However, wild-type mice showed a well described dose-dependent opiate-induced increase in locomotion, whereas *NK1*^{−/−} mice showed no increase in activity even at high doses of drug, although pre-drug levels of activity were again comparable (Fig. 1c). Thus, unlike dopamine D2 receptor knockout mice¹⁰, the *NK1*^{−/−} mutant mice were not hypokinetic and locomotion did not increase with opiate administration. Intermediate doses of morphine (6–10 mg kg^{−1}) significantly reduced locomotion of *NK1*^{−/−} mice but this returned to control levels at higher doses (20 mg kg^{−1}) of drug. These results strongly indicate that the locomotor effects of morphine require release of substance P.

The locomotor-stimulating action of morphine may be linked to the rewarding properties of the drug¹². We investigated the motivational salience of morphine using a conditioned place preference (CPP) procedure. Mice learn to associate drug with one of two compartments that are visually and texturally dissimilar. During the preconditioning phase, both *NK1*^{−/−} and wild-type mice visited the compartments of the place-preference apparatus, spending equivalent amounts of time in each compartment. Morphine administered subcutaneously at a dose of 1 mg kg^{−1} did not induce a place preference in either genotype. However, at 3 mg kg^{−1} a strong preference emerged in wild-type mice but not in *NK1*^{−/−} mice. Wild-type animals preferred the compartment that had

become associated with morphine administration. Saline administered to control mice of both genotypes had no effect on place preference (Fig. 2). This dose of morphine produced no drug-induced changes in locomotor activity in either genotype.

We next investigated whether this deficit in the rewarding effect of morphine was also true of other rewards such as the psychotropic drug cocaine or palatable food. Using the same CPP procedure, mice of either genotype spent proportionately more time in the compartment associated with 5 mg kg^{−1} cocaine, administered by intraperitoneal injection (Fig. 2). A lower dose of cocaine (2.5 mg kg^{−1}) reduced CPP (by 34% in wild-type and 38% in *NK1*^{−/−} mice) but it was significantly different from saline-injected mice and there was no difference between genotypes. CPP at higher doses of cocaine (10 mg kg^{−1}) was indistinguishable from that seen at 5 mg kg^{−1}, indicating a ceiling effect. The rewarding effects of cocaine may be mediated by brain mechanisms partially distinct from those involved in opiate reward^{5–8}; our data would support this contention. Similarly, when food was used as the reward, place

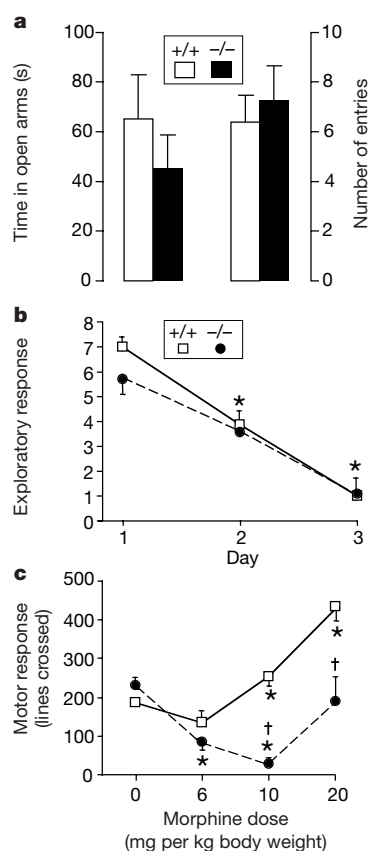


Figure 1 Locomotor behaviour does not increase in *NK1*^{−/−} mice after morphine administration. **a**, Behaviour of the *NK1*^{−/−} mice in the elevated plus maze. Time (s) spent in the open arms and number of visits to the open arm were similar in both genotypes in control. Data are mean $\pm$ s.e.m. and were analysed by two-way analysis of variance and Dunnett's test, $n = 10$ per group. **b**, Exploratory response of *NK1*^{−/−} and wild-type mice in a stressful open field containing unknown objects. Values represent the number of visits to the unknown objects (mean $\pm$ s.e.m.). Number of squares crossed for wild-type were: day 1, 116.2 $\pm$ 20.9; day 2, 89.7 $\pm$ 13.9; day 3, 80 $\pm$ 26; and for *NK1*^{−/−} were: day 1, 111.6 $\pm$ 10.4; day 2, 113 $\pm$ 7.2; day 3, 123 $\pm$ 13. Number of rearings and defecation was similar for the three days and between genotypes. Asterisk, $P < 0.05$ versus same group on day 1 (Dunnett's test, $n = 10$). **c**, Motor response to morphine administration in wild-type (squares) and *NK1*^{−/−} (circles) mice, in a non-stressful environment. Number of line crossings were measured after injection of different doses of morphine. Values represent mean $\pm$ s.e.m. ($n = 6$). Asterisk, $P < 0.05$ versus saline group with the same genotype; cross, $P < 0.001$ versus wild-type receiving the same treatment (two-tailed Student's t -test).

preference developed in both wild and $NK1^{-/-}$ mice. This was independent of the motivational state of the animal, as preference developed in both food-deprived and sated mice (Fig. 3).

Having established that the $NK1^{-/-}$ mice had a specific deficit in the opiate reward mechanism, we investigated whether this generalized to a failure to register negative reinforcement¹³. That is, would mice dependent on opiates avoid a compartment associated with naloxone-precipitated withdrawal? Opioid dependence was induced in mice by repeated intraperitoneal injection of morphine twice daily over a period of five days, with a gradually increasing dose (from 10 to 50 mg kg⁻¹). The mice were maintained at 50 mg kg⁻¹ morphine for a further two days. The conditioning phase consisted of two consecutive days of first naloxone and then vehicle injection. Wild-type mice showed a clear aversion to the compartment associated with withdrawal, whereas the $NK1^{-/-}$ mice showed significantly less avoidance of that compartment (Fig. 4). These data indicate that $NK1^{-/-}$ mice are impaired in detecting both the rewards of morphine and its withdrawal.

Mice also develop a physical dependence on repeated administration of opiates. We investigated the somatic signs of naloxone-mediated withdrawal following dependence induced by twice daily

injections of morphine increasing from 20 to 100 mg kg⁻¹. As before, injections of saline were used in control groups of $NK1^{-/-}$ and wild-type mice. Both groups of mice showed the Straub reflex, a classical sign of exposure to opiates. Somatic signs of withdrawal were evaluated immediately after naloxone injection for a period of 30 min (Table 1 in Supplementary information). Withdrawal signs were not present before naloxone treatment. Both wild-type and knockout mice displayed the full range of classical withdrawal symptoms^{12,14} with the exception of jumping behaviour, which was absent in $NK1^{-/-}$ mice (scores: $NK1^{+/+}$ 23.9 ± 6.2 and $NK1^{-/-}$ 0.8 ± 0.5, $P < 0.05$). Jumping behaviour is regarded as one of the 'dominant' motor signs of withdrawal¹⁴. This blunting of the withdrawal response is consistent with the previously reported effects of intracerebrally injected NK1 antagonists on naloxone-induced withdrawal in rats¹⁴.

The clear disruption of opiate rewarding behaviour seen in the $NK1^{-/-}$ mice could be due to a secondary loss of mu opiate receptor activity, which is known to initiate this and to mediate analgesia¹⁵. However, a general loss of opiate receptor activity appeared unlikely for two reasons. First, the knockout mice show most opiate withdrawal signs; and second, the central analgesic actions of morphine are present and not attenuated in $NK1^{-/-}$ mice tested on a hot-plate task¹. Compensatory developmental changes could have occurred in the knockout mouse that would account for the behavioural phenotype. However, the differences in behaviour seen in $NK1^{-/-}$ mice have been replicated in other species using specific NK1 antagonists (see Supplementary Information and refs 1, 16).

Areas of the forebrain distinct from those that generate dependence can mediate the rewarding effects of opiates^{17,18}. We therefore examined the striatal distribution and binding kinetics of the mu receptor and second messenger (cyclic AMP) response to morphine in receptors isolated from striatal membranes. Immunohistochemistry showed that the mu receptor localized in a similar 'patch-matrix' fashion¹⁹ within the striatum and nucleus accumbens and at comparable levels of immunoreactivity in both wild-type and mutant mice. Saturation binding with the selective mu receptor ligand [³H]DAMGO ((D-Ala², N-MePhe⁴, Gly-ol⁵)-enkephalin) to forebrain membrane preparations²⁰ indicated that wild-type and mutant mice have comparable numbers of receptors and receptor affinity. We measured a maximal binding (B_{max}) of 0.16 ± 0.037 pmol μ g⁻¹ protein and a dissociation constant (K_d) of 6.12 ± 1.73 nM for wild-type mice, and a B_{max} of 0.12 ± 0.01 pmol μ g⁻¹ protein and K_d of 4.4 ± 0.5 nM for $NK1^{-/-}$ mice. Finally, cAMP production following mu receptor stimulation²¹ by an acute morphine challenge was assayed in crude striatal homogenates as a

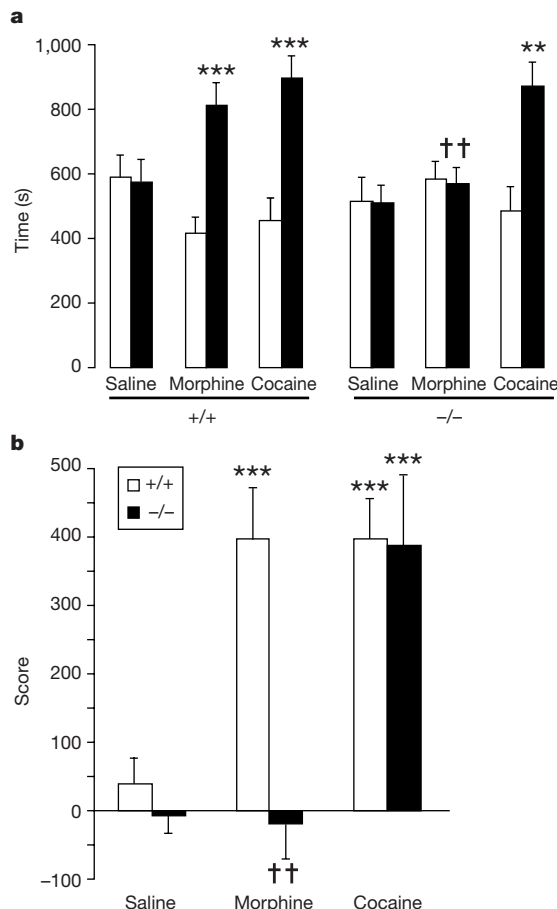


Figure 2 Lack of conditioned place preference induced by morphine in $NK1^{-/-}$ mice. **a**, Time spent in the drug-associated compartment during the preconditioning (white bars) and the testing phases (black bars). **b**, Scores calculated as the difference between post-conditioning and pre-conditioning time spent in the compartment associated with the drugs. Morphine treatment was subcutaneous injection of 3 mg kg⁻¹; cocaine treatment was intraperitoneal injection of 5 mg kg⁻¹; saline was injected as control. Data were analysed with Dunnett's test ($n = 12$ for the morphine group, $n = 10$ for saline group, $n = 8$ for the cocaine group). Values represent mean ± s.e.m.; triple asterisk, $P < 0.001$; double asterisk, $P < 0.02$ versus saline group with the same genotype; double cross, $P < 0.02$, versus wild-type group receiving the same treatment.

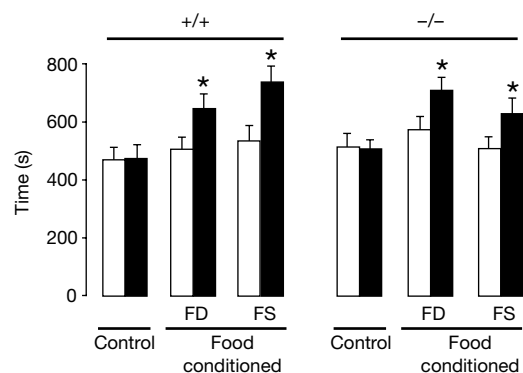


Figure 3 Conditioned place preference induced by food reward in $NK1^{-/-}$ mice and wild-type litter mates in food-deprived (FD) and food-sated (FS) animals. Time spent in the food-associated compartment during the preconditioning (white bars) and testing (black bars) phases in control and in food-conditioned group. Asterisk, $P < 0.05$ versus the non-conditioned group with the same genotype (two-tailed Student's t -test, $n = 8$).

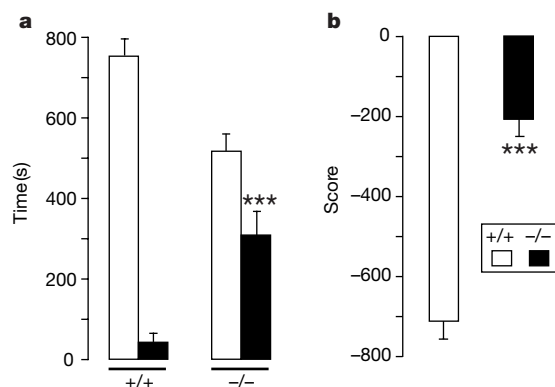


Figure 4 Reduced conditioned response when place is paired with naloxone-induced withdrawal in $NK1^{-/-}$ mice. **a**, Time spent in the drug-associated compartment during the preconditioning (white bars) and the testing phases (black bars). **b**, Scores calculated as the difference between postconditioning and preconditioning time spent in the

compartment associated with the drugs. Naloxone was injected intraperitoneally at 1 mg kg^{-1} . Values represent mean $\pm$ s.e.m., triple asterisk, $P < 0.001$ versus wild-type group receiving the same treatment (two-tailed Student's t -test; $n = 6$ for wild-type group, $n = 10$ for $NK1^{-/-}$ group).

measure of adenylyl cyclase activity. The percentage reduction of cAMP production following acute opiate treatment was 43% and 60% for the wild-type and mutant mice, respectively. In summary, the mu opiate receptor was functional in $NK1^{-/-}$ mice.

Because of the role of the dopamine receptor in opiate-mediated reward¹⁰, we also examined the expression of the D1 and D2 receptors using *in situ* hybridization. Again there were no differences in the forebrain expression of these two genes between $NK1^{-/-}$ and wild-type mice, indicating that dopamine receptor expression was unaltered in the mutant mouse.

Finally, the long-term plastic changes that develop following opiate or psycho-stimulant treatment are linked to the expression of immediate early genes, particularly Fos-B and the truncated Δ Fos-B isoform within the nucleus accumbens²². We examined the expression of Fos-B using semi-quantitative immuno-fluorescence and confocal microscopy. We found no significant difference in the total number of Fos-B-positive neurons in the nucleus accumbens between wild-type and $NK1$ knockout mice following chronic morphine treatment. However, although there was a substantial increase in the fluorescence intensity of Fos-B positive neurons in both 'core' and 'shell' zones of the nucleus accumbens of wild-type mice compared with saline controls (core 46%; $P < 0.01$ and shell 35%; $P < 0.05$), no such increase was seen in $NK1^{-/-}$ mice. This result indicates that Fos-B activation may be linked to the expression of opiate-dependent reward^{22,23}.

Our results provide compelling evidence for a role of substance P in opiate reward mechanisms and in addiction. The data argue strongly that a key synapse in this process lies within the ventral forebrain and occurs between collaterals of substance-P-releasing populations of striatal projection neurons and large cholinergic neurons of the nucleus accumbens and nucleus basalis that express the NK1 receptor²⁴. These neurons are implicated in associated learning and respond to stimuli that trigger a learned and rewarded task^{25–27}.

Stress is a precipitating factor in causing relapse into drug taking in man and drug self-administration in animals²⁸. However, stress responses can be attenuated by substance P receptor antagonists³ or by genetic disruption of the substance P receptor^{1,16}. Therefore, drugs that antagonize the actions of substance P may be powerful new tools in both the treatment of opiate drug addiction and the prevention of relapse into drug taking. □

Methods

Gene targeting

$NK1^{-/-}$ mice and $NK1^{+/+}$ littermates were derived as described¹ from the mating of heterozygous $NK1^{+/-}$ mice. The targeting construct was derived from a mouse 129/sv

strain genomic library and targeted clones were injected into C57BL/6 blastocysts. Chimaeric males were mated with C57BL/6 females. Mice were bred from successive generations of sibling knockout and wild-type mice and can be thought of as representing a recombinant inbred strain. For CPP we also tested mice that were pure C57/B6 or 129/sv and these behaved exactly as wild-type mice (data not included).

Locomotor assay

The behaviour of mice was videotaped in an open field and observations were made 10 min after injection of drug or saline¹. Animals were habituated to the experimental room for 24 h but were not habituated to the open field. Illumination was set at 40 lux.

Elevated plus maze

This assay is based on the natural aversion of rodents to height and open spaces. The plus maze apparatus was made of black plexiglas and had two open arms ($30 \times 5 \text{ cm}$) and two enclosed arms of the same size with transparent walls 30 cm high (to obtain homogeneous illumination in the whole apparatus, 6 lux). It was elevated 30 cm above ground. The arms were connected by a central square platform ($5 \times 5 \text{ cm}$) and the maze formed a cross. Male mice were habituated to the experimental room for 24 h and the subjects were individually tested in 5-min sessions. Each animal was placed on the centre platform facing an open arm to initiate the test session and recorded with a video camera. Behaviours scored were number of open arm and closed arm entries and time spent on the various sections of the maze.

Exploratory response

This was examined in a white open field ($50 \times 50 \times 50 \text{ cm}$) that was brightly illuminated (500 lux) and contained four unknown objects, placed 20 cm from each corner. This test was performed on three consecutive days, at 24-h intervals¹².

Place preference

The apparatus and conditioning procedure has been described¹². Analysis of CPP without food deprivation was assessed in animals with limited access to food (15% of mouse body as food with sucrose added) for 6 days before the test. Conditioning phase was performed with access to food on days 1, 3 and 5 and with no access to food on days 2, 4 and 6. Control animals had no access to food in the assigned compartments during this phase. In the food-deprived condition, mice were deprived of food the nights before testing.

Aversive place conditioning

See ref. 15. Morphine dependence was induced with twice daily intraperitoneal administration of morphine. The doses of morphine injected at 9:00 and 19:00 on consecutive days were: day 1, 10 mg kg^{-1} ; day 2, 20 mg kg^{-1} ; day 3, 30 mg kg^{-1} ; day 4, 40 mg kg^{-1} ; and days 5–7, 50 mg kg^{-1} . The conditioning phase consisted of two consecutive days of alternative naloxone or vehicle injection. One hour after the morning injection of morphine, animals were injected intraperitoneally with naloxone (day 1, 1 mg kg^{-1}) or saline (day 2) and immediately confined for 20 min in the appropriate compartment.

Adenylyl cyclase activity

Mice treated acutely with morphine or saline were killed 1 h later. Tissue was homogenized; homogenates incubated with ATP and the amount of cAMP formed was determined with a [^3H] cAMP protein binding assay²³ (Amersham).

Binding to brain homogenates

Binding was performed using 100–150 μg total brain membrane proteins with

[³H]DAMGO (Amersham)¹⁸. Measurements were done in triplicate. Binding data was analysed using the LIGAND program (supplied by P. Munson).

In situ hybridization

Cryostat sections (13 µm) of fresh brain tissue were hybridized with ³⁵S-labelled oligonucleotide probes specific for the dopamine D1 and D2 receptors¹ (see Supplementary Information).

Immunohistochemistry

Mice were anaesthetized with sodium pentobarbital and fixed transcardially with 4% paraformaldehyde in 0.1 M phosphate buffer, the brains removed and sections cut at 40 µm for processing with antibodies against the mu opiate receptor (Chemicon), c-Fos or Fos-B (Santa Cruz). The Fos-B antibody recognizes both full-length and truncated products of the gene. To study Fos-B expression, mice were treated daily with two injections of morphine in increasing doses (20, 40, 60, 80 mg kg⁻¹). On the last day, a single dose of 100 mg kg⁻¹ was given in the morning and mice were perfused 6 h later.

Received 30 September 1999; accepted 2 March 2000.

- De Felipe, C. *et al.* Altered nociception, analgesia and aggression in mice lacking the receptor for substance P. *Nature* **392**, 394–397 (1998).
- Kramer, M. S. *et al.* Distinct mechanism for antidepressant activity by blockade of central substance P receptors. *Science* **281**, 1640–1645 (1998).
- Culman, J., Klee, S., Ohlendorf, C. & Unger, T. Effect of tachykinin receptor inhibition in the brain on cardiovascular and behavioural responses to stress. *J. Pharmacol. Exp. Ther.* **280**, 238–246 (1997).
- Nakaya, Y. *et al.* Immunohistochemical localization of substance P receptor in the central nervous system of the adult rat. *J. Comp. Neurol.* **347**, 249–274 (1994).
- Wise, R. A. Neurobiology of addiction. *Curr. Opin. Neurobiol.* **6**, 243–251 (1996).
- Robbins, T. W. & Everitt, B. J. Drug addiction: bad habits add up. *Nature* **398**, 567–570 (1999).
- Koob, G. F., Sanna, P. P. & Bloom, F. E. Neuroscience of addiction. *Neuron* **21**, 467–476 (1998).
- Koob, G. K. & Le Moal, M. Drug abuse: Hedonic homeostatic dysregulation. *Science* **278**, 52–70 (1997).
- Gerfen, C. R. Substance P (neurokinin-I) receptor mRNA is selectively expressed in cholinergic neurons in the striatum and basal forebrain. *Brain Res.* **556**, 165–170 (1991).
- Maldonado, R. *et al.* Absence of opiate rewarding effects in mice lacking dopamine D2 receptors. *Nature* **388**, 586–589 (1995).
- Baik, J. H. *et al.* Parkinsonian-like locomotor impairment in mice lacking dopamine D2 receptors. *Nature* **377**, 424–428 (1995).
- Wise, R. A. & Bozarth, M. A. A psychomotor stimulant theory of addiction. *Psychol. Rev.* **97**, 469–492 (1987).
- Bechara, A. & van der Kooy, D. A single brain stem substrate mediates the motivational effects of both opiates and food in nondeprived rats but not in deprived rats. *Behav. Neurosci.* **106**, 351–363 (1992).
- Maldonado, R., Girdlestone, D. & Roques, B. P. RP 67580, a selective antagonist of neurokinin-1 receptors, modifies some of the naloxone-precipitated morphine withdrawal signs in rats. *Neurosci. Lett.* **156**, 135–140 (1993).
- Matthes, H. W. D. *et al.* Loss of morphine-induced analgesia, reward effect and withdrawal symptoms in mice lacking the m-opioid-receptor gene. *Nature* **383**, 819–823 (1996).
- Rupniak, N. M. J. *et al.* A Pharmacological blockade or genetic deletion of substance P (NK1) receptors attenuates neonatal vocalisation in guinea-pigs and mice. *Neuropharmacology* (in press).
- Maldonado, R., Stimus, L., Gold, L. & Koob, G. F. Role of different brain structures in the expression of the physical morphine withdrawal syndrome. *J. Pharmacol. Exp. Ther.* **261**, 669–677 (1992).
- Bozarth, M. A. Physical dependence produced by central morphine infusions: an anatomical mapping study. *Neurosci. Biobehav. Rev.* **18**, 373–383 (1994).
- Mansour, A., Fox, C. A., Burke, S., Akil, H. & Watson, S. J. Immunohistochemical localization of the cloned mu opiate receptor in the rat CNS. *J. Chem. Neuroanat.* **8**, 283–305 (1995).
- Kieffer, B. L., Beford, K., Gaveriaux-Ruff, C. & Hirth, C. G. The delta-opioid receptor: isolation of a cDNA by expression cloning and pharmacological characterization. *Proc. Natl Acad. Sci. USA* **89**, 12048–12052 (1992).
- Izenwasser, S., Buzas, B. & Cox, B. M. Differential regulation of adenylyl cyclase activity by mu and delta opioids in rat caudate putamen and nucleus accumbens. *J. Pharmacol. Exp. Ther.* **267**, 145–152 (1993).
- Nye, H. E. & Nestler, E. J. Induction of chronic Fos-related antigens in rat brain by chronic morphine administration. *Mol. Pharmacol.* **49**, 636–645 (1996).
- Kelz, M. B. *et al.* Expression of the transcription factor delta FosB in the brain controls sensitivity to cocaine. *Nature* **401**, 272–276 (1999).
- Aubry, J. M., Schulz, M. F., Pagliusi, S., Schulz, P. & Kiss, J. Z. Coexpression of dopamine D2 and substance P (neurokinin-1) receptor messenger RNAs by a subpopulation of cholinergic neurons in the rat striatum. *Neuroscience* **53**, 417–424 (1993).
- Graybiel, A. M., Aosaki, T., Flaherty, A. W. & Kimura, M. The basal ganglia and adaptive motor control. *Science* **265**, 1826–1831 (1994).
- Elliott, P. J. & Iversen, S. D. Behavioural effects of tachykinins and related peptides. *Brain Res.* **381**, 68–76 (1986).
- Boix, F., Sandor, P., Nogueira, P. J., Huston, J. P. & Schwarting, R. K. Relationship between dopamine release in nucleus accumbens and place preference induced by substance P injected into the nucleus basalis magnocellularis region. *Neuroscience* **64**, 1054–1055 (1995).
- Shaham, Y. & Stewart, J. Stress reinstates heroin-seeking in drug-free animals: an effect mimicking heroin, not withdrawal. *Psychopharmacol.* **119**, 334–341 (1995).

Supplementary information is available on Nature's World Wide Web site (<http://www.nature.com>) or as a paper copy from the London editorial office of Nature.

Acknowledgements

We thank J. O'Brien, J. A. Perez De Gracia, and P. Mantyh, R. Maldonado and C. Stanford

for reading and commenting on the manuscript, and N. Rupniak for sharing unpublished data.

Correspondence and requests for materials should be sent to S.P.H. (e-mail: hunt@ucl.ac.uk).

Vanilloid receptor-1 is essential for inflammatory thermal hyperalgesia

John B. Davis*, Julie Gray*, Martin J. Gunthorpe*, Jonathan P. Hatcher*, Phil T. Davey*, Philip Overend†, Mark H. Harries*, Judi Latcham‡, Colin Clapham§, Kirsty Atkinson§, Stephen A. Hughes§, Kim Rance§, Evelyn Grau§, Alex J. Harper*, Perdita L. Pugh*, Derek C. Rogers*, Sharon Bingham*, Andrew Randall* & Steven A. Sheardown§

Departments of *Neuroscience Research, †Statistical Sciences, ‡Laboratory Animal Sciences and §Biotechnology & Genetics, SmithKline Beecham Pharmaceuticals, Third Avenue, Harlow CM19 5AW, UK

The vanilloid receptor-1 (VR1) is a ligand-gated, non-selective cation channel expressed predominantly by sensory neurons. VR1 responds to noxious stimuli including capsaicin, the pungent component of chilli peppers, heat and extracellular acidification, and it is able to integrate simultaneous exposure to these stimuli^{1,2}. These findings and research linking capsaicin with nociceptive behaviours (that is, responses to painful stimuli in animals³) have led to VR1 being considered as important for pain sensation. Here we have disrupted the mouse VR1 gene using standard gene targeting techniques. Small diameter dorsal root ganglion neurons isolated from VR1-null mice lacked many of the capsaicin-, acid- and heat-gated responses that have been previously well characterized in small diameter dorsal root ganglion neurons from various species. Furthermore, although the VR1-null mice appeared normal in a wide range of behavioural tests, including responses to acute noxious thermal stimuli, their ability to develop carrageenan-induced thermal hyperalgesia was completely absent. We conclude that VR1 is required for inflammatory sensitization to noxious thermal stimuli but also that alternative mechanisms are sufficient for normal sensation of noxious heat.

We have used homologous recombination in embryonic stem (ES) cells to generate a mouse lacking transmembrane domains 2–4 of the mVR1 gene (Fig. 1a). Germline chimaeras were crossed onto C57Bl/6J females to generate heterozygotes, which were intercrossed giving rise to overtly healthy homozygous mutant offspring in the expected mendelian ratio (average litter numbers: VR1^{-/-}, 2.5 ± 0.36; VR1^{+/-}, 4.9 ± 0.40; VR1^{+/+}, 2.5 ± 0.23; n = 28). Successful targeting of the locus and germline transmission was confirmed by Southern analysis (Fig. 1b) and by polymerase chain reaction (PCR) (Fig. 1c).

The expression of VR1 in sensory neurons from dorsal root ganglia has been previously established using both messenger RNA analysis and immunocytochemistry^{1,2}. In addition, numerous pharmacological assays using vanilloid receptor agonists, such as capsaicin and resiniferatoxin, and antagonists, such as capsazepine⁴, have demonstrated the presence of functional VR1 in small diameter dorsal root ganglion (sDRG) neurons⁵. We have used a similar functional approach to confirm the elimination of VR1-mediated responses in our VR1-null mice. Dissociated neuronal cultures were prepared from the DRG of VR1 wild type (VR1^{+/+}), VR1 heterozygote (VR1^{+/-}) or VR1-null (VR1^{-/-}) littermates. After 20–40 h in culture, electrophysiological recordings of responses to a range of stimuli were made from the classically nociceptive, small DRG neurons (diameter <22 µm, mean capacitance 9.4 ± 0.5 pF).

Responses to capsaicin (1 μ M) were present in 74% ($n = 34$) of sDRG neurons from VR1^{+/+} mice. A similar number (62%, $n = 21$; data not shown) from VR1^{+/-} mice also responded to capsaicin, with currents indistinguishable in size or waveform from those observed in VR1^{+/+} cells. In contrast, capsaicin failed to elicit a response in any of the 29 cells tested from VR1^{-/-} cultures (Fig. 2).

In 9 out of 13 cells from VR1^{+/+} cultures a transient extracellular acidification to pH 5.3 produced a slowly developing, non-desensitizing, inward current. This current had kinetics that closely resembled the slow acid-gated current observed in primary rat DRG neurons⁶, or in HEK 293 cells or *Xenopus* oocytes expressing VR1 (ref. 1). This slow acid-gated current was not observed in any cells that did not respond to capsaicin. In 46% of cells from VR1^{+/+} cultures, however, an additional rapidly activating and desensitizing inward current was also observed at the onset of a response to pH 5.3. This transient current component was found in both capsaicin-responsive and -unresponsive cells and probably represents the activation of channels belonging to the acid-sensing ionic channel (ASIC) family⁷. In all VR1^{-/-} cells examined, the slowly activating pH 5.3-gated response was absent. In contrast, the rapidly activating ASIC-type current was still seen in a proportion of cells (see Fig. 2a). These data support the hypothesis that a principal part of the maintained (that is, non-desensitizing) acid-gated excitations of sDRG neurons reflects the activity of VR1. This finding also indicates that VR1 may be important in generating acidosis-related pain.

Although lacking the expected responses to pH 5.3 and capsaicin, VR1^{-/-} cells responded normally to two neurotransmitters, GABA (γ -aminobutyric acid, 100 μ M) and ATP (50 μ M), that have been previously shown to activate ion channels in sensory neurons^{8,9}. The

responses to these agents were present in a similar percentage of VR1^{+/+} and VR1^{-/-} cells (Fig. 2b). Furthermore, no marked differences in response amplitudes or kinetics were noted when recordings from VR1^{+/+} and VR1^{-/-} cells were compared. Notably, the responses to ATP in both VR1^{+/+} and VR1^{-/-} sDRG neurons exhibited the rapid desensitization kinetics typical of capsaicin-responsive sDRG neurons⁹.

Heterologously expressed recombinant VR1, in addition to responding to capsaicin, acidic pH and anandamide, is activated by increases in temperature to levels above $\sim 44^\circ\text{C}$ (refs 1, 10). Similar heat-activated currents have been reported in capsaicin-responsive sDRG neurons^{11,12}. To test the contribution of VR1 to these thermal responses, we compared the heat-gated currents generated by 2-s step increases in temperature, from room temperature (23.5–25.5 $^\circ\text{C}$) up to 54 $^\circ\text{C}$, in VR1^{+/+} and VR1^{-/-} sDRG neurons. As reported¹¹, the heat-activated current exhibits two components (Fig. 3a, top two panels). The first component shares characteristics with currents mediated by recombinant VR1, including a well defined activation threshold at $\sim 44^\circ\text{C}$ and substantial outward rectification. The second has been ascribed to a 'non-specific' change in membrane or seal resistance¹¹. This latter current does not exhibit a marked activation threshold¹¹, is associated with little or no change in current variance and has an amplitude directly proportional to the pre-stimulus holding current (and input resistance). Such non-specific heat-gated currents can also be observed in capsaicin-unresponsive sDRG neurons as well as in other cell types.

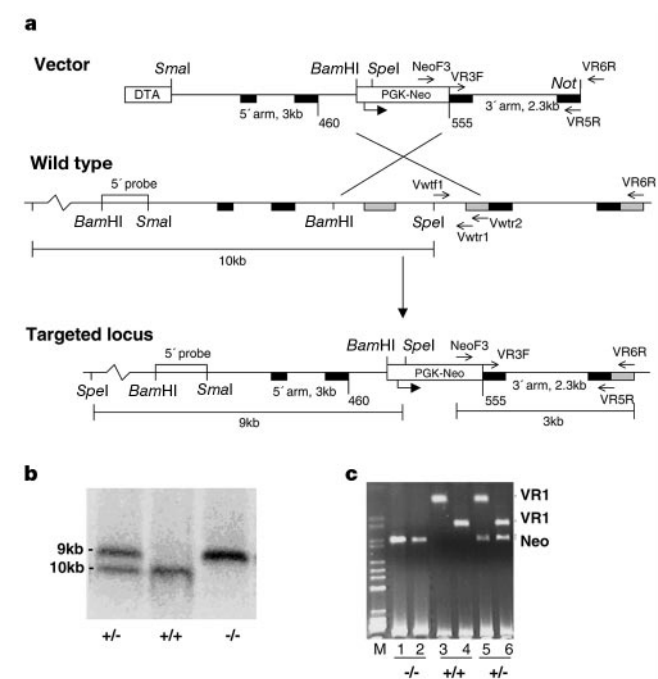


Figure 1 Disruption of VR1 gene by homologous recombination. **a**, Diagram of targeting construct and strategy. Exons are indicated as black boxes. Shaded boxes represent exon sequences not present in the targeting construct. The diagnostic PCR product and *SpeI* cleavage fragments are indicated with the 3' external PCR primer VR6R and the 5' external *BamHI/SmaI* probe (open box). **b**, Representative example of genomic Southern blot from wild type (+/+), heterozygous mutant (+/-) and homozygous mutant (-/-) mouse tail DNA cleaved with *SpeI* and hybridized with a ³²P-labelled *BamHI/SmaI* 5' external probe. **c**, Wild type (+/+), heterozygous mutant (+/-) and homozygous mutant (-/-) mouse tail DNA genotyped by multiplex PCR. All samples contained neo-specific primers. Lanes 1, 3 and 5 contain wild-type VR1-specific primers Vwtr1 and Vwtr2. Lanes 2, 4 and 6 contain wild-type VR1-specific primers Vwtr1 and Vwtr2.

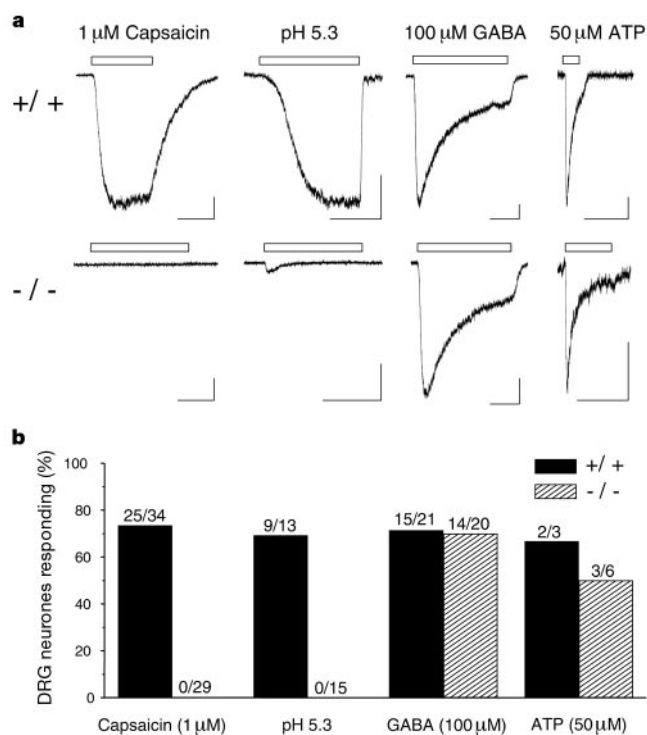


Figure 2 Ligand-gated currents in sDRG neurons from VR1^{+/+} and VR1^{-/-} mice. **a**, Whole-cell patch-clamp recordings of inward currents recorded in sDRG neurons, from VR1^{+/+} or VR1^{-/-} mice, in response to the application of capsaicin (1 μ M), protons (extracellular acidification to pH 5.3), GABA (100 μ M) or ATP (50 μ M) for the time indicated by the bar. Vertical scale bar represents 200 pA for currents recorded in response to capsaicin, pH 5.3 and GABA, and 50 pA for ATP-gated current. The horizontal scale bar represent 5 s in all cases. **b**, Pooled data from all VR1^{+/+} (black bars) and VR1^{-/-} (hatched bars) DRG neurons studied indicating the percentage of cells responsive to the application of either capsaicin, GABA or ATP, or yielding a sustained response to protons.

Increases in temperature (from 25 to 54 °C) in both VR1^{+/+} and VR1^{-/-} sDRG neurons consistently produced a heat-activated inward current. In a proportion of VR1^{+/+} cells, the current elicited had an amplitude that was linearly related to the applied temperature (Fig. 3a, top) and was thus attributable to the non-specific current. In the remaining VR1^{+/+} sDRG neurons, a clear additional current component was recruited at temperatures above a threshold of ~44 °C (Fig. 3a, middle, and 3b). The proportion of VR1^{+/+} cells exhibiting this additional heat-gated current (15 out of 33) was similar to the proportion of capsaicin-responsive cells (5 out of 11) in the same preparation (~50%). Pooled data comparing the heat-gated responses at 49–50 °C in VR1^{+/+} with VR1^{-/-} sDRG neurons show that no VR1^{-/-} cells responded with a response greater than 200% of the holding current at room temperature (mean 79 ± 10%), whereas this level was surpassed in almost 50% of VR1^{+/+} cells (Fig. 3c). The lack of a 'specific' current in VR1^{-/-} sDRG neurons strongly supports previous suggestions that a current with its characteristics arises from the thermal activation of VR1 (ref. 12).

On the basis of its heat-dependent gating, VR1 has been proposed to be a sensor for noxious heat¹. To specifically address this hypothesis, we compared the thermoresponsive behaviour of VR1^{+/+}, VR1^{+/-} and VR1^{-/-} mice, both acutely and after the experimental induction of inflammation. The latency of response to an acute thermal stimulus was measured using a 50 or 52.5 °C hotplate (Fig. 4a), or a focused radiant heat stimulus (Fig. 4c,d). All of the mice studied responded to these previously well characterized noxious stimuli.

Increasing the temperature of the hotplate decreased the latency of response for all genotypes (Fig. 4a, $F = 24.91$; d.f. = 1,49; $P < 0.001$). Unexpectedly, neither VR1^{-/-} mice ($P = 0.16$ and $P = 0.06$, 50 and 52.5 °C, respectively) nor VR1^{+/-} mice ($P = 0.52$ and $P = 0.08$, 50 and 52.5 °C, respectively) showed a significant difference when compared with VR1^{+/+} mice in hotplate responses (Fig. 4a). Similarly, there were no significant differences between genotypes in

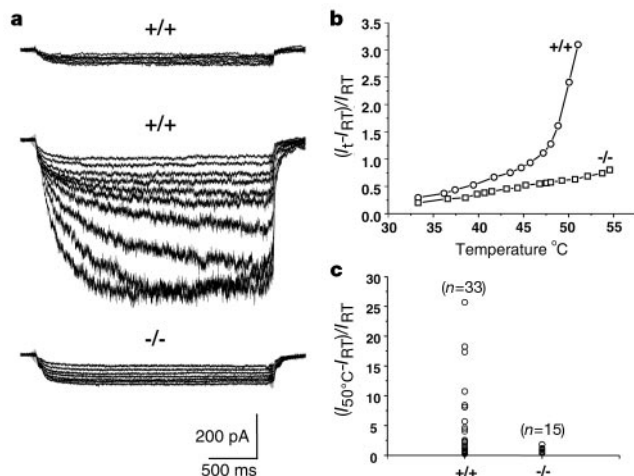


Figure 3 Heat-gated currents in sDRG neurons from VR1^{+/+} and VR1^{-/-} mice. **a**, Example of heat-activated currents from three different sDRGs. Top traces, a VR1^{+/+} cell lacking a detectable threshold-dependent heat-activated current (temperature range 33–54 °C); middle traces, a VR1^{+/+} cell exhibiting a clear threshold-gated current (temperatures 33.3, 37.4, 41.7, 44.8, 45.7, 47.2, 48.8, 50.1, 51 °C); bottom trace, a typical VR1^{-/-} cell (temperature range 33–54 °C). **b**, Plot of applied-heat versus evoked current for the responsive VR1^{+/+} and VR1^{-/-} cells shown in **a**. The peak amplitude for each temperature has been normalized to the pre-stimulus room-temperature (RT) holding current. **c**, Scatter plot showing the increase in current amplitude produced by a temperature jump to 49–50 °C. Data points are normalized to the pre-stimulus room-temperature holding current. Note the significant population of VR1^{+/+} cells exhibiting a heat-gated response greater than 200% of holding current. These cells all possessed a slowly activating high-variance current akin to that shown in the middle panel of **a**.

terms of response to a radiant heat stimulus set at an intensity that produces a paw-withdrawal latency of ~9 s in normal mice (Fig. 4c). However, the analysis of variance of the effect of genotype upon response latency at 52.5 °C ($F = 3.11$; d.f. = 2,49 $P = 0.053$) only just failed to reach significance. Thus, although it is clear that mice lacking VR1 are still able to respond to noxious heat, there remains a possibility that subtle differences in the behavioural responses to different intensities of heat may occur. Such an effect may be masked to some degree by the mixed strain background of the N1F1 cohort that we used in these studies.

Our data give rise to a paradox because although the electrophysiological response specific to noxious temperatures was absent in the VR1^{-/-} sDRG neurons the null mice responded normally to noxious heat. Neonatal treatment with capsaicin, which ablates the capsaicin-sensitive nociceptors, results in significant analgesia towards thermal stimuli¹³, suggesting that the VR1-expressing nociceptors are essential for normal thermal nociception. Future work may help to explain the paradox, for example, by characterization of the DRG cell types responsible for transducing the retained thermal sensation or by the discovery of another heat sensor whose function may be lost or compromised in cell culture. Vanilloid-receptor like 1 (VRL-1), which has also been reported as a heat receptor¹⁴, may contribute to the retained response. VRL-1 has a higher threshold for activation (~52 °C compared with 42 °C for VR1 *in vitro*)¹⁴ and may therefore be unlikely to have a significant

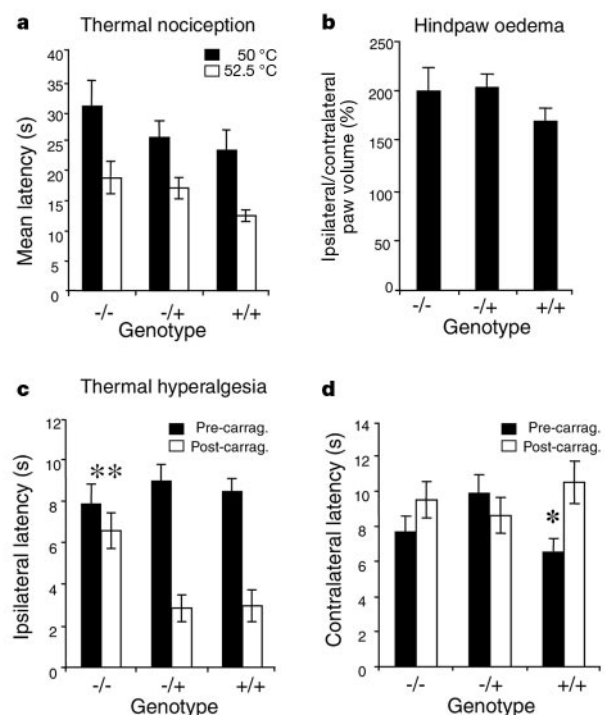


Figure 4 Responses to thermal stimuli in normal and sensitized animals. **a**, Normal nociceptive behaviour to noxious thermal stimuli. Withdrawal or licking latencies in the hotplate assay at 50 or 52.5 °C. No significant differences were seen between wild-type (+/+; $n = 8$ and 9 for 50 °C and 52.5 °C, respectively), heterozygous (+/-; $n = 10$ and 10 , respectively), or homozygous mutant (-/-; $n = 10$ and 9 , respectively) mice. **b**, Measurement of paw volume after injection of carrageenan. Ipsilateral volume is expressed as a percentage of the uninjected contralateral paw ($n = 10$ for each group). **c**, Ipsilateral withdrawal latencies before or 4 h after carrageenan injection, in response to focused radiant heat. VR1^{-/-} mice showed a highly significant reversal of hyperalgesia when compared with VR1^{+/+} mice; two asterisks, $P < 0.001$. **d**, Contralateral withdrawal latencies before or after carrageenan injection, in response to radiant heat. VR1^{+/+} ($n = 17$) but not VR1^{-/-} ($n = 18$) or VR1^{+/-} ($n = 20$) mice show a significant hypoalgesic effect; asterisk, $P < 0.05$.

role in the heat responses described above. Developmental adaptation in the VR1-null mice, or uncharacterized properties of VRL-1 *in vivo*, however, can not be ruled out at this stage. Single or double genetic knockout approaches may shed further light on the relative contributions of VR1, VRL-1 or other molecular heat detectors.

A hyperalgesic response to thermal stimuli is associated with inflammation and was tested for using intraplantar carrageenan injection into the hindpaw and subsequent stimulation by radiant heat¹⁵. Analysis of the degree of carrageenan-induced hindpaw inflammation showed no significant differences in swelling between VR1^{-/-} and VR1^{+/+} or VR1^{-/+} animals (Fig. 4b). Four hours after carrageenan injection, VR1^{+/+} mice and VR1^{-/+} mice exhibited a highly significant decrease in paw-withdrawal latencies compared with baseline pre-inflammation responses (Scheffé's test, $P < 0.001$ and $P < 0.001$, respectively). A markedly different result was found in VR1^{-/-} mice. In these animals, withdrawal latencies of inflamed paws were indistinguishable from those measured before carrageenan injection (Scheffé's test, $P = 0.955$, Fig. 4c). Thus, knock out of VR1 eliminated the generation of a thermal hyperalgesic response. The contralateral paw showed no significant changes in response latency for the VR1^{-/-} and VR1^{-/+} mice, but there was an increase compared with baseline values in contralateral latencies in VR1^{+/+} mice. This may reflect a chance finding or the onset of an endogenous inhibitory drive resulting in hypoalgesia in the uninflamed contralateral paw (Fig. 4d).

These data highlight a marked effect of VR1 deletion upon responsiveness to noxious thermal stimuli where nociceptive pathways are sensitized (that is, during inflammation), and contrast with the situation in the naive mouse where thermal nociception is relatively unchanged. Pharmacological experiments using the VR1 antagonist capsazepine have also demonstrated an antihyperalgesic effect¹⁶. Selective VR1 antagonists may therefore prove effective for the treatment of thermal hyperalgesia.

Immunohistochemical and PCR analysis has identified VR1 expression in a wide range of brain regions and has led to suggestions that VR1 might be involved in behaviours other than nociception^{3,17}. To screen for further behavioural phenotypes and examine the specificity of our observations, VR1^{-/-} mice were tested for overt behavioural phenotypes using the 'SHIRPA' tests¹⁸. This assessment included 38 behavioural observations, recording of spontaneous locomotor activity and a holeboard exploratory behaviour test; the data arising from the screen are available in the Supplementary Information. Also included in this screen was a test of mechanical nociception—a toe pinch test. No significant differences between VR1^{+/+} and VR1^{-/-} mice were observed in any of these tests. These negative data demonstrate the specificity of the phenotype resulting from VR1 disruption, and also exclude altered nociceptive responses occurring as a result of effects upon motor function.

Our data demonstrate the specificity of the VR1 knockout in terms of gross behavioural deficits and point to an essential or predominant role of VR1 in responses of sDRG neurons to capsaicin, extracellular acidification or heat. Observations of the activation of VR1 by these and other ligands, such as anandamide¹⁰, suggest that VR1 may also function in pain where these other mediators are the predominant agonists. Our experiments have focused upon thermal responses, however, and the results show conclusively that VR1 is essential for the development of sensitization to thermal stimuli during inflammation but not for the normal sensation of noxious heat. □

Methods

Targeting of VR1 gene and generation of mutant mice

A standard gene targeting approach was chosen to replace, in E14.1 ES cells, the DNA encoding amino acids 460–555 of mVR1 with the neomycin phosphotransferase gene (detailed methods are provided in Supplementary Information). A 2.3-kilobase (kb) 3'

and 3.75-kb 5' homology arm were isolated from a mouse 129SVJ BAC library, cloned either side of PGKneo, and flanked by Diphtheria toxin-A gene as a negative selection marker¹⁹. Homologous recombination in resistant ES cells was confirmed by Southern blot and a multiplex PCR genotyping procedure^{20,21}. Three targeted clones were injected into C57Bl6/J-derived blastocysts, and the resulting chimaeras produced germline offspring. Males heterozygous for the mutated allele were mated to C57Bl6/J females and mutant progeny were intercrossed to generate an N1F1 study population. All experiments were conducted according to the requirements of the United Kingdom Animals (Scientific Procedures) Act (1986) and conformed to the ethical standards of SmithKline Beecham Pharmaceuticals.

Electrophysiological recordings

Dissociated DRG cells were prepared from 8–10-day-old pups and cultured on glass coverslips³ in DMEM with N2 supplements, 50 ng ml⁻¹ β -NGF, 0.05% bovine serum albumin, 100 U ml⁻¹ penicillin and streptomycin. Whole-cell patch-clamp recordings were made from sDRGs ($\leq 22 \mu\text{M}$ diameter) with an Axopatch 200B amplifier controlled by the pClamp7 software suite (Axon Instruments). For analysis of ligand- and pH-gated current, all experiments were conducted at room temperature (23–25.5 °C) at a holding potential of -70 mV. The bath solution comprised (in mM) NaCl 130, KCl 5, CaCl₂ 2, MgCl₂ 1, glucose 30, HEPES-NaOH 25; pH 7.3. Corresponding pH 5.3 solutions were adjusted with HCl. To minimize the holding current, and thus the contribution of the 'non-specific' heat-gated currents, most of our recordings of heat-gated current were made in near symmetrical Cs⁺ concentrations, with extracellular Ba²⁺ substituted for Ca²⁺ and methanesulphonate as the predominant intracellular anion. However, essentially similar observations (data not shown) were also made in a population of recordings under the conditions above. Electrodes had a resistance of 2–7 M Ω when filled with (in mM): CsCl 140, MgCl₂ 4, EGTA 10, HEPES-CsOH 10; pH 7.3. Applications of agonists and room temperature to 54 °C temperature jumps were made using an automated fast-switching solution exchange system (Warner Instruments SF-77B; time for solution exchange ~30 ms), in a manner similar to that described elsewhere¹.

Behavioural studies

All tests were performed on male 6–9-week-old mice ($n = 10$). VR1^{-/-} and VR1^{+/+} littermates were used in the primary behavioural observation screen SHIRPA¹⁸, full experimental procedures for which are available at <http://www.mgu.har.mrc.ac.uk/mutabase/>. In addition, a 30-min spontaneous locomotor activity test and a 10-min holeboard test of exploratory activity were used²².

Hotplate test

Mice of each genotype ($n = 8–10$) were tested in a random and blind fashion for thermal nociception using a hotplate (Harvard Analgesia Meter, Harvard Instruments) maintained at 50 or 52.5 °C. Mice were observed for signs of nociception, that is, rapid fanning or licking of the paws. The response latency was recorded and results analysed using 2-way analysis in Statistica (Statsoft Inc.). Genotype and hotplate temperature were used as independent variables. Follow up analyses were carried out using Duncan's test.

Carrageenan-induced inflammation

The same cohort of mice ($n = 17–20$ per genotype) were tested for thermal hyperalgesia using described methods modified for mice¹⁴. Animals were habituated to the test apparatus. Baseline withdrawal latencies, to a focused radiant heat stimulus (Plantar Test Apparatus, Ugo Basile), were measured for each hindfoot. Carrageenan (0.025 mls, 2%) was injected sub-plantar into the left hindfoot, and paw-withdrawal latencies were measured for both hindfeet 4 h later. Paw volumes were measured using a plethysmometer (Linton Instruments). Results were expressed as mean latency (s) for paw withdrawal at baseline and after carrageenan injection. All data were log transformed to correct for heterogeneity of variances. Results were analysed in SAS (Statsoft Inc.) using a split plot analysis of variance. Genotype and measurement were used as independent variables. In both cases, follow up analysis was carried out using Scheffé's test, where appropriate²³. Errors shown represent standard errors of the mean.

Received 20 December 1999; accepted 14 April 2000.

1. Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
2. Tominaga, M. *et al.* The cloned capsaicin receptor integrates multiple pain-producing stimuli. *Neuron* **21**, 531–543 (1998).
3. Szallasi, A. & Blumberg, P. M. Vanilloid (Capsaicin) receptors and mechanisms. *Pharmacol. Rev.* **51**, 159–211 (1999).
4. Bevan, S. *et al.* Capsazepine: A competitive antagonist of the sensory neurone excitant capsaicin. *Br. J. Pharmacol.* **107**, 544–552 (1992).
5. Wood, J. N. *et al.* Capsaicin-induced ion fluxes in dorsal root ganglion cells in culture. *J. Neurosci.* **8**, 3208–3220 (1988).
6. Bevan, S. & Yeats, J. Protons activate a cation conductance in a sub-population of rat dorsal root ganglion neurones. *J. Physiol.* **433**, 145–161 (1991).
7. Waldmann, R. *et al.* H⁺-gated cation channels. *Ann. NY Acad. Sci.* **868**, 67–76 (1999).
8. Bevan, S. & Winter, J. Nerve growth factor (NGF) differentially regulates the chemosensitivity of adult rat cultured sensory neurones. *J. Neurosci.* **15**, 4918–4926 (1995).
9. Ueno, S., Tsuda, M., Iwanaga, T. & Inoue, K. Cell type-specific ATP-activated responses in rat dorsal root ganglion neurones. *Br. J. Pharmacol.* **126**, 429–436 (1999).
10. Zygmunt, P. M. *et al.* Vanilloid receptors on sensory nerves mediate the vasodilatory action of anandamide. *Nature* **400**, 452–457 (1999).

11. Cesare, P. & McNaughton, P. A novel heat-activated current in nociceptive neurons and its sensitization by bradykinin. *Proc. Natl Acad. Sci. USA* **93**, 15435–15439 (1996).
12. Nagy, I. & Rang, H. P. Similarities and differences between the responses of rat sensory neurons to noxious heat and capsaicin. *J. Neurosci.* **19**, 10647–10655 (1999).
13. Nagy, J. I. & Kooy, D. van der. Effects of neonatal capsaicin treatment on nociceptive thresholds in the rat. *J. Neurosci.* **3**, 1145–1150 (1983).
14. Caterina, M. J., Rosen, T. A., Tominaga, M., Brake, A. J. & Julius, D. A capsaicin-receptor homologue with a high threshold for noxious heat. *Nature* **398**, 436–441 (1999).
15. Hargreaves, K., Dubner, R., Brown, F., Flores, C. & Joris, J. A new and sensitive method for measuring thermal nociception in cutaneous hyperalgesia. *Pain* **32**, 77–88 (1988).
16. Kwak, J. Y., Jung, J. Y., Hwang, S. W., Lee, W. T. & Oh, U. A capsaicin-receptor antagonist, capsazepine, reduces inflammation-induced hyperalgesic responses in the rat: evidence for an endogenous capsaicin-like substance. *Neuroscience* **86**, 619–626 (1998).
17. Sasamura, T., Sasaki, M., Tohda, C. & Kuraishi, Y. Existence of capsaicin-sensitive glutamatergic terminals in rat hypothalamus. *Neuroreport* **9**, 2045–2048 (1998).
18. Rogers, D. C. *et al.* 'SHIRPA'—a comprehensive behavioural and functional analysis of mouse phenotype. *Mamm. Genome* **8**, 711–713 (1997).
19. Yagi, T. *et al.* A novel selection for homologous recombinants using diphtheria toxin A fragment gene. *Anal. Biochem.* **214**, 77–86 (1993).
20. Torres, R. M. & Kuhn, R. *Laboratory Protocols for Conditional Gene Targeting*. (Oxford Univ. Press, Oxford, 1997).
21. Hooper, M., Hardy, K., Handyside, A., Hunter, S. & Monk, M. HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature* **326**, 292–295 (1987).
22. Rogers, D. C. *et al.* Use of SHIRPA and discriminant analysis to characterise marked differences in the behavioural phenotype of six inbred mouse strains. *Behav. Brain Res.* **105**, 207–217 (1999).
23. Miliken, G. A. & Johnson, D. E. in *Analysis of Messy Data* 29–45 (Chapman & Hall, London, 1992).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

The authors would like to acknowledge P. Hayes, J. Nation, S. Pickering and C. David for technical assistance, and S. Rastan, F. Walsh, M. Geppert and D. Simmons for valuable critique.

Correspondence or requests for materials should be addressed to J.B.D. (e-mail: John_B_Davis@sphrd.com).

Glutamatergic synapses on oligodendrocyte precursor cells in the hippocampus

Dwight E. Bergles*, J. David B. Roberts†, Peter Somogyi† & Craig E. Jahr*

* Vollum Institute, L474, Oregon Health Sciences University, Portland, Oregon 97201, USA

† MRC Anatomical Neuropharmacology Unit, Department of Pharmacology, University of Oxford, Oxford OX1 3TH, UK

Fast excitatory neurotransmission in the central nervous system occurs at specialized synaptic junctions between neurons, where a high concentration of glutamate directly activates receptor channels. Low-affinity AMPA (α -amino-3-hydroxy-5-methyl isoxazole propionic acid) and kainate glutamate receptors are also expressed by some glial cells¹, including oligodendrocyte precursor cells (OPCs). However, the conditions that result in activation of glutamate receptors on these non-neuronal cells are not known. Here we report that stimulation of excitatory axons in the hippocampus elicits inward currents in OPCs that are mediated by AMPA receptors. The quantal nature of these responses and their rapid kinetics indicate that they are produced by the exocytosis of vesicles filled with glutamate directly opposite these receptors. Some of these AMPA receptors are permeable to calcium ions, providing a link between axonal activity and internal calcium levels in OPCs. Electron microscopic analysis revealed that vesicle-filled axon terminals make synaptic junctions with the

processes of OPCs in both the young and adult hippocampus. These results demonstrate the existence of a rapid signalling pathway from pyramidal neurons to OPCs in the mammalian hippocampus that is mediated by excitatory, glutamatergic synapses.

Oligodendrocytes in the mammalian central nervous system develop from a population of precursor cells during late gestational and early postnatal life², providing the insulating sheaths of myelin necessary for rapid conduction of action potentials along axons. These precursors or OPCs were identified anatomically as smooth protoplasmic astrocytes based on their unique stellate morphology³, and their properties have been studied both in culture (termed O-2A cells)^{4–6} and in acutely isolated tissue (termed glial precursors or complex cells)^{1,7}. Glutamate receptor activation in these cells inhibits their proliferation and maturation into oligodendrocytes⁸, and prolonged exposure to glutamate causes excitotoxic degeneration⁹. Despite the potential importance of this pathway in the development and regeneration of myelin, it is not known how these receptors are activated *in vivo*. Glutamate has been shown to reach other glial cells by diffusion from nearby synaptic clefts following vesicular release¹⁰, or by reverse transport along axons¹¹.

To determine how glutamate reaches AMPA receptors on OPCs, we made whole-cell patch-clamp recordings from OPCs in the hippocampus and measured their response to stimulation of afferent excitatory axons. OPCs located in the stratum radiatum region of area CA1 exhibited small Na^+ currents, large A-type and delayed rectifier K^+ currents, and did not fire action potentials (Fig. 1a) ($n = 28$). Electrical stimulation in stratum radiatum elicited inward currents in OPCs that had rapid kinetics (Fig. 1b). Cells with these properties had a stellate morphology, with thin, highly branched processes that extended from a small cell body (Fig. 1c). They were identified as OPCs by their immunoreactivity to NG2 (Fig. 1d, e, $n = 10/10$), a proteoglycan that is only expressed by OPCs in this region¹². These NG2-positive cells were immunonegative for glial fibrillary acidic protein ($n = 3/3$), while astrocytes recorded from under similar conditions were immunopositive ($n = 8/8$ groups of cells).

Paired stimuli produced currents in OPCs that were larger ($\text{P2/P1} = 1.7 \pm 0.1$, $n = 16$) and exhibited fewer apparent failures following the second stimulus (Fig. 2a), similar to paired-pulse facilitation of excitatory postsynaptic currents (EPSCs) in CA1

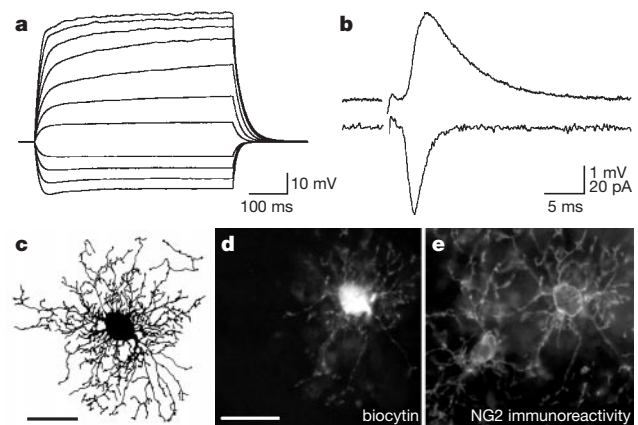


Figure 1 Synaptic responses from identified OPCs in hippocampal slices. **a**, Current-clamp recording of membrane responses to current injection (-80 to 140 pA; step size, 20 pA). **b**, Evoked responses to Schaffer collateral/commissural fibre stimulation, recorded in voltage-clamp (holding potential, -90 mV; lower trace) and current-clamp (membrane potential = -90 mV; upper trace). Traces are averages of 15 consecutive responses recorded from the same cell. Stimulus: 30 μA , 100 μs . **c**, Reconstruction of a biocytin-filled OPC. **d**, Micrograph of the same OPC as in **d**, visualized by AMCA-conjugated streptavidin. **e**, NG2-immunoreactivity of same region of the slice as shown in **d**. Scale bars for **c** and **d**, 20 μm ; **d** and **e** are at the same magnification.

pyramidal neurons. These currents were blocked by the competitive AMPA/kainate receptor antagonist 2,3-dihydroxy-6-nitro-7-sulphamoyl-benzo(F)quinoxaline (NBQX, 5 μ M; $n = 28$) (Fig. 2a), and strongly inhibited (>95%) by 1-(4-aminophenyl)-4-methyl-7,8-methylenedioxy-5H-2,3-benzodiazepine hydrochloride (GYKI 52466, 50 μ M), a non-competitive AMPA receptor antagonist ($n = 4$). Cyclothiazide (CTZ, 100 μ M), which potentiates currents mediated by AMPA, but not kainate receptors¹³, increased the amplitude ($282 \pm 41\%$, $n = 12$) and slowed the time constant of decay ($\tau_{\text{control}} = 1.21 \pm 0.09$ ms, $\tau_{\text{CTZ}} = 3.76 \pm 0.14$ ms, $n = 12$) of evoked responses. The sensitivity of these currents to both CTZ and GYKI 52466 indicates that they are mediated by AMPA receptors. Evoked currents in OPCs were also reversibly blocked by Cd^{2+} (30 μ M), indicating that they are dependent on Ca^{2+} release mechanisms and are not produced by reverse transport of

glutamate^{11,14} (Fig. 2b). EPSCs in CA1 pyramidal neurons are markedly enhanced by antagonists of A_1 adenosine receptors, such as 8-cyclopentyltheophylline (CPT), resulting from the relief of presynaptic inhibition by ambient adenosine¹⁵. The amplitudes of OPC responses were also increased ($252 \pm 41\%$, $n = 6$) by CPT (1 μ M), indicating that the release of glutamate onto OPCs is under similar inhibitory control (Fig. 2b).

AMPA receptors lacking the edited GluRB (GluR2) subunit are permeable to Ca^{2+} and exhibit rectifying current-voltage (I - V) relationships due to channel block by internal polyamines¹⁶. To determine whether AMPA receptors activated in OPCs with axonal stimulation were permeable to Ca^{2+} , we examined the I - V relationship of evoked responses. Slight inward rectification was observed under control conditions that was enhanced by including spermine (25 μ M) in the internal solution (Fig. 2c), indicating that Ca^{2+} -permeable AMPA receptors contribute to the evoked synaptic currents. In addition, these currents were inhibited by Joro spider toxin (5 μ M, $n = 4$), a selective antagonist of Ca^{2+} -permeable AMPA/kainate receptors¹³. These results indicate that activity of glutamatergic afferents is linked to Ca^{2+} entry into OPCs.

Axons from CA3 pyramidal neurons form synapses with dendritic spines of CA1 pyramidal neurons in the stratum radiatum region¹⁶. To test whether excitatory inputs to OPCs also arise from CA3 pyramidal neurons, we recorded from OPCs in stratum radiatum of area CA1 and focally applied L-glutamate (200 μ M) onto the cell body layer of area CA3 (Fig. 2d). In several cells (3/11), glutamate application elicited brief bursts of inward currents in

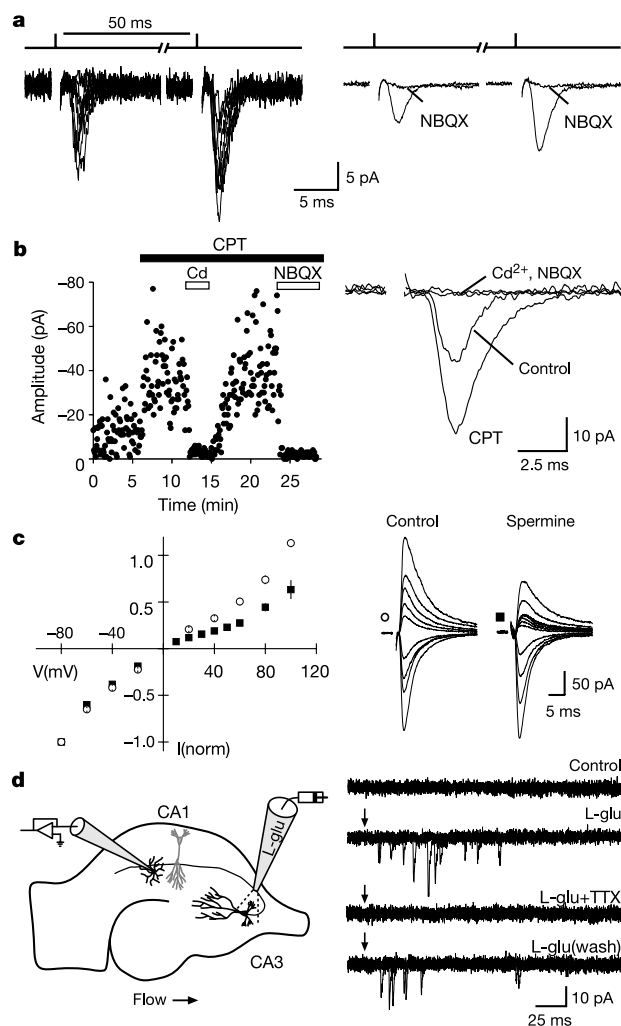


Figure 2 Properties of excitatory responses in OPCs. **a**, Facilitation of evoked responses by paired stimuli. Evoked inward currents were blocked by NBQX (5 μ M). Traces at right are averages of 15 consecutive responses. **b**, Plot of peak amplitudes of evoked responses that were increased by CPT (1 μ M) and blocked by cadmium 'Cd' (CdCl_2 , 30 μ M) or NBQX (5 μ M). Traces at right are averages of 15 consecutive responses. **c**, I - V relationship of evoked responses under control conditions (open circles), or with spermine (25 μ M) in the intracellular solution (filled squares). ACSF contained CTZ (100 μ M) and CPT (1 μ M). Traces at right are averages of 10 consecutive responses recorded at -80 to 100 mV (step size, 20 mV). **d**, Puffer application of L-glutamate (200 μ M, 50 ms; down arrow) onto CA3 pyramidal neurons (diagram at left) elicited bursts of inward currents in OPCs that were blocked by TTX (1 μ M). Traces at right are overlays of 5 consecutive sweeps.

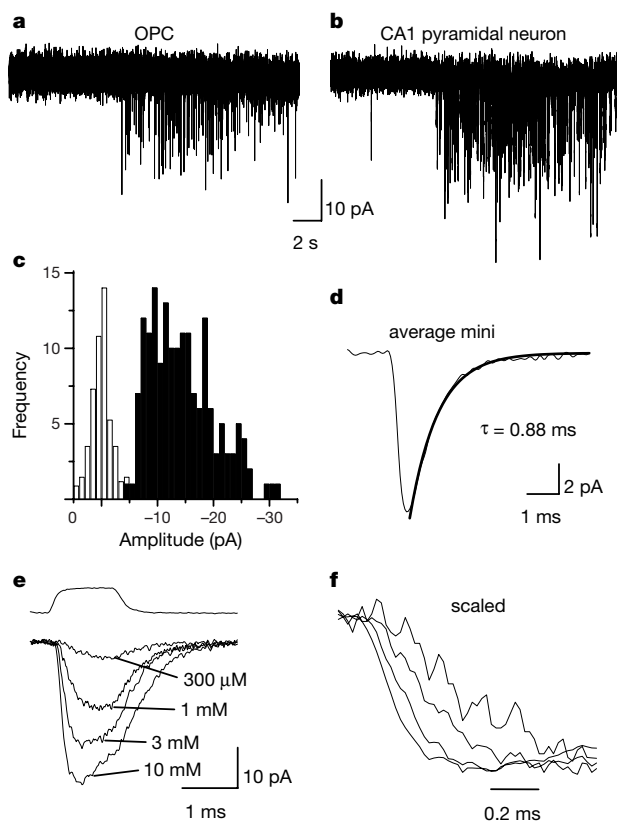


Figure 3 Quantal EPSCs in OPCs are elicited by a brief, high concentration of glutamate. Pardaxin (2 μ M) triggered bursts of mEPSCs in OPCs (**a**) and CA1 pyramidal neurons (**b**). **c**, Amplitude distribution of mEPSCs recorded from an OPC. **d**, Average time course of 169 mEPSCs recorded from an OPC. Thick line is a single exponential fit to the decay. **a**, **c**, **d** are from the same cell. **e**, Concentration dependence of the kinetics of AMPA/kainate receptors in outside-out patches from OPCs. Traces are averages of 30 consecutive responses. **f**, The rising phase of responses in **e** shown scaled to the peak of the response to 10 mM L-glutamate.

OPCs (Fig. 2d). These glutamate-evoked currents were blocked by the Na^+ channel antagonist, tetrodotoxin (TTX, $1 \mu\text{M}$), indicating that they were dependent on propagation of action potentials from CA3. Because the only glutamatergic neurons located in the CA3 region are pyramidal neurons¹⁶, these results suggest that Schaffer collateral axons of CA3 pyramidal neurons provide excitatory input to both OPCs and CA1 pyramidal neurons.

Spontaneous fusion of transmitter-filled vesicles occurs at individual excitatory synapses in the absence of action potentials,

resulting in miniature EPSCs (mEPSCs)¹⁷. Similar miniature AMPA receptor-mediated currents were observed in OPCs at a very low frequency ($<0.01 \text{ Hz}$). Brief application of the presynaptic neurotoxin picrotoxin ($2 \mu\text{M}$), which enhances the frequency of vesicular release from nerve terminals¹⁸, caused the appearance of high-frequency bursts of mEPSCs in both OPCs ($n = 11$ cells) (Fig. 3a) and CA1 pyramidal neurons ($n = 6$ cells) (Fig. 3b); similar bursts were induced in OPCs by α -latrotoxin (2 nM , $n = 4$ cells). The amplitudes of mEPSCs recorded from OPCs were highly variable

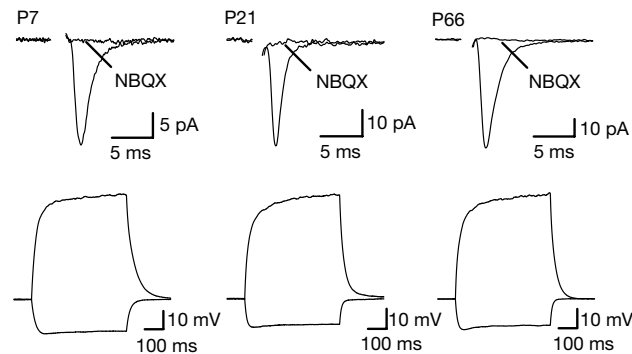


Figure 4 Excitatory synaptic responses were evoked in OPCs from both young and adult animals. EPSCs from OPCs in hippocampal slices prepared from different age animals were blocked by NBQX ($5 \mu\text{M}$). Stimulus intensity was $20\text{--}30 \mu\text{A}$, $100 \mu\text{s}$. ACSF

contained CPT ($1 \mu\text{M}$). Lower traces show membrane responses of OPCs from these age animals to current injection (-60 and 120 pA).

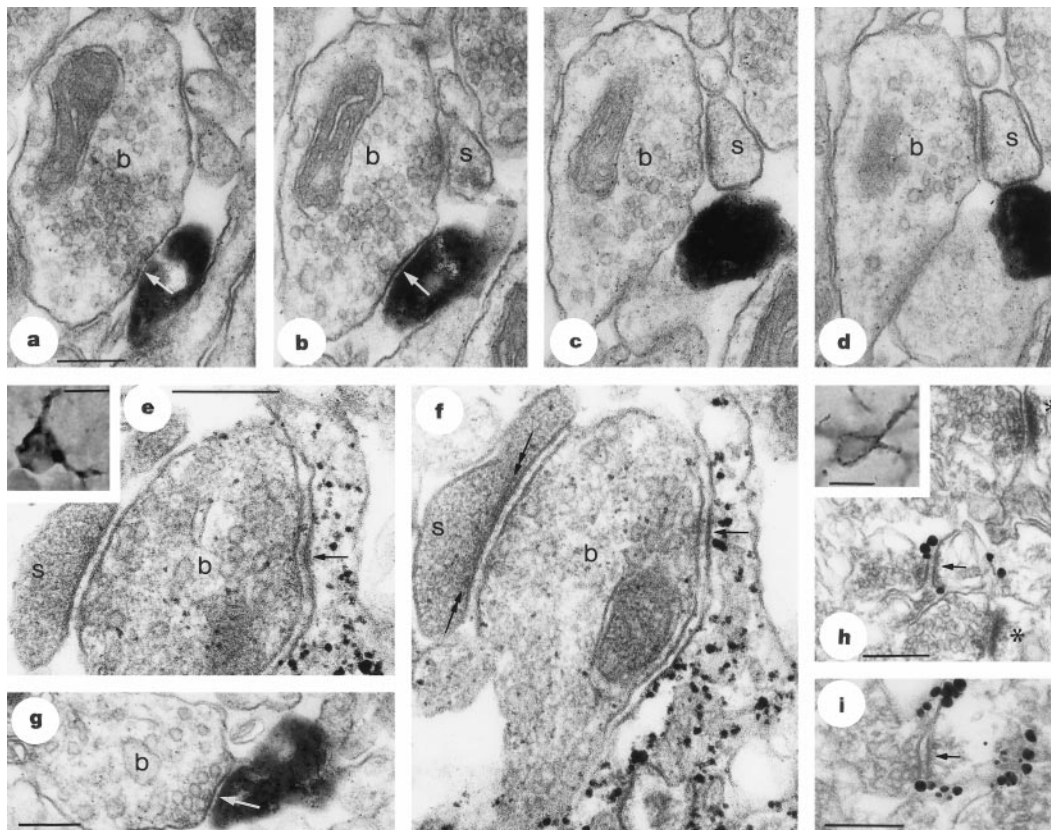


Figure 5 Electron micrographs of the synaptic relationships of OPCs. **a–d**, Serial sections of a process (black, peroxidase reaction) from a physiologically identified, biocytin labelled OPC receiving a synapse (arrow) from a bouton (b) that also gives a synapse to a dendritic spine (s). The OPC process approaches the spine postsynaptic density to within $\sim 110 \text{ nm}$ (**c**). **e, f**, Silver-intensified gold reaction of an OPC (inset, light micrograph) process reveals the thinner postsynaptic membrane specialization (arrows) as compared with the postsynaptic density (between double arrows) of a spine (s) innervated by the same bouton

(**b**). **g**, Another bouton (b) made a synapse (arrow) only with an OPC process. **h, i**, NG2-immunopositive OPC (inset) visualized with an antibody to an extracellular epitope, resulting in silver/gold particles (electron-opaque dots) surrounding OPC processes in stratum radiatum of adult rat. The postsynaptic membrane specialization is thinner than that of neighbouring synapses on dendritic spines (asterisks). Scale bars for **a–i**, $0.2 \mu\text{m}$; scale bars in insets, $10 \mu\text{m}$.

(coefficient of variance, $35 \pm 4\%$; mean amplitude, 15 ± 3 pA; $n = 11$ cells), and did not conform to a normal distribution (Fig. 3c), similar to mEPSCs recorded from neurons in many brain regions¹⁷. The small size of the mEPSCs and the indication that they arise from Schaffer collateral fibres that have a low release probability¹⁹, suggests that approximately 60 inputs would have to be stimulated to elicit a -100 pA synaptic current in an OPC. The average time course of these mEPSCs was very fast, rising in 259 ± 15 μ s (20–80%; range, 179–335 μ s; $n = 11$ cells) and decaying with a time constant of 983 ± 58 μ s (range, 700–1,250 μ s; $n = 11$ cells) (Fig. 3d).

To estimate the amount of glutamate necessary to produce AMPA receptor-mediated mEPSCs in OPCs, we measured the concentration-dependence of the kinetics of these receptors in isolation, by applying different concentrations of glutamate to outside-out patches removed from OPCs. A concentration of 3 mM L-glutamate was necessary to produce currents with a rise time similar to the mEPSCs (10 mM, 180 ± 29 μ s; 3 mM, 286 ± 48 μ s; 1 mM, 397 ± 56 μ s; 300 μ M, $1,184 \pm 191$ μ s; $n = 5$) (Fig. 3e, f). These data and the less than threefold increase in OPC synaptic responses produced by CTZ²⁰ indicates that the mEPSCs are not produced by a low concentration glutamate diffusing out of nearby neuronal synapses¹⁰. These results suggest that numerous excitatory synapses are formed on OPCs.

The link between glutamate receptor activation and OPC differentiation⁸ raised the possibility that these excitatory synapses are present only during early development. However, OPCs with similar electrophysiological properties were identified in slices prepared from both young (postnatal day 7, P7, $n = 4$; P21, $n = 10$), and mature animals (P66, $n = 5$) (Fig. 4), consistent with previous immunocytochemical results indicating that these cells are abundant in both developing and adult brain^{7,12}, and OPCs from all ages received similar excitatory synaptic input (Fig. 4).

Biocytin loading of OPCs made it possible to identify the relationships of their processes to neuronal synaptic membranes. Electron microscopy revealed presynaptic transmitter release sites that were identified on the basis of clustering of synaptic vesicles to a membrane region exhibiting electron-opaque clumps of granular material (presynaptic dense projections) (Fig. 5). In the first type of synaptic relationship, OPC processes were directly apposed to presynaptic release sites as seen for cells visualized by peroxidase ($n = 2$) (Fig. 5a–d, g) or gold/silver ($n = 2$) methods (Fig. 5e, f). Axonal and glial membranes were rigidly apposed in the synaptic junction (49 synapses, 7–19 per cell, 44 completely serially sectioned), and contained electron-opaque cleft material and an intracleft disk. Four presynaptic terminals also made another synaptic junction with a nearby dendritic spine in the sectioned zone (Fig. 5b–g), confirming the common origin of innervating afferents from CA3. The postsynaptic membrane specialization of the glial membrane was much thinner than that of nearby type 1 synapses on spines (Fig. 5e, f, h, i), and most synaptic junctions received by OPCs were small (largest extent: 0.211 ± 0.077 μ m (s.d.); range, 0.062–0.430 μ m). The existence of direct synaptic junctions on OPCs suggests that they are responsible for the fast synaptic currents.

In the second type of synaptic relationship, the processes of OPCs were closely aligned to the synaptic cleft of axo-spinous synaptic junctions (Fig. 5c, h). The closest distance between the edge of the postsynaptic density and the OPC process was 74 ± 6 nm (range, 38–136 nm; $n = 22$; P21 rat). In a series of sections, 22 close alignments to spine synapses and 19 direct synapses were found over 101 μ m² of OPC membrane, representing about 2.5% of the cell. If cells *in vivo* receive a similar innervation, this suggests that an OPC can receive about 800 direct synapses.

The perisynaptic location of some OPC processes is similar to astrocytic²¹ and Bergmann glial cell processes³. Glutamate concentration peaks at less than 200 μ M at Bergmann glial membranes following release at climbing fibre–Purkinje neuron synapses in the

cerebellum^{20,22}, producing slow AMPA receptor-mediated currents in these cells (20–80% rise time, 1 ms; half decay time, 4.2 ms)²². The fast kinetics of OPC responses make it unlikely that they are produced by spillover of glutamate from nearby neuronal synapses. However, the frequent close opposition of OPC processes to axo-spinous synapses raises the possibility that these cells modulate neuronal synaptic transmission.

It is possible that synapses on OPCs are formed only following slice preparation. Immunogold/silver labelling of OPCs *in situ*, for an extracellular epitope of the NG2 proteoglycan in the CA1 hippocampal area from perfusion fixed adult and developing rats, revealed direct synaptic junctions (Fig. 5h, i) ($n = 24$; 22 completely sectioned) with similar characteristics to those observed in acutely isolated tissue. These results indicate that neuron–OPC synapses are present *in vivo*.

Oligodendrocyte development is regulated by secreted factors⁷ and direct cell–cell interactions²³. This evidence of functional glutamatergic synapses between CA3 pyramidal neurons and CA1 OPCs in the hippocampus during the period of oligodendrocyte maturation², provides a pathway for axons to regulate myelination. In the adult central nervous system, OPCs remain numerous in the grey matter where myelination is sparse, and they do not readily proliferate to replace oligodendrocytes that have been damaged or lost through lesions²⁴ or demyelinating diseases such as multiple sclerosis²⁵; this suggests that OPCs are not maintained solely as a reservoir for further oligodendrocyte production. The persistence of a rapid excitatory signalling pathway between axons and OPCs raises the possibility of an additional neuromodulatory role for these cells, possibly through the release of neuroactive substances in response to glutamatergic synaptic input. □

Methods

Slide preparation

Hippocampal slices were prepared from male Sprague–Dawley rats (12–16 d old) in accordance with a protocol approved by the Department of Animal Care at OHSU as described²⁶, and maintained in a solution consisting of (in mM): 119 NaCl, 2.5 KCl, 2.5 CaCl₂, 1.3 MgCl₂, 1 NaH₂PO₄, 26.2 NaHCO₃, and 11 glucose, saturated with 95% O₂/5% CO₂.

Electrophysiology

Putative OPCs located in stratum radiatum of area CA1 were visualized with infrared/Nomarski videomicroscopy. Cells were selected for recording that had a small round cell body (soma diameter <10 μ m), with no large processes extending from it. The internal solution for current-clamp recordings contained (in mM): 130 K-methanesulphonate, 20 HEPES, 10 EGTA, 2 Mg-ATP, 0.4 Na-GTP; pH 7.3, and for voltage-clamp recordings contained (in mM): 100 Cs-methanesulphonate, 20 TEA-Cl, 20 HEPES, 10 EGTA, 1 MgCl₂, 4 Mg-ATP, 0.3 Na-GTP; pH 7.3. Holding potentials have been adjusted for a 10-mV junction potential (K-methanesulphonate, 10.2 mV; Cs-methanesulphonate, 10.4 mV). The input resistance of OPCs was 257 ± 28 M Ω ($n = 23$) with K-methanesulphonate, and >1 G Ω with Cs-methanesulphonate.

Evoked responses were elicited with a constant-current stimulator (Winston Electronics) using bipolar stainless-steel electrodes (tip separation ~ 200 μ m) placed in stratum radiatum >50 μ m from the cell (stimulus: 10–40 μ A, 100 μ s). Current-clamp recordings were made using an Axoclamp-2A amplifier (Axon Instruments). Whole-cell evoked currents were recorded at -90 mV using an Axopatch 200A (Axon Instruments), filtered at 2 kHz and sampled at 10–50 kHz. Exogenous L-glutamate (200 μ M) was ejected from a patch pipette to stratum pyramidale of area CA3 with pulses of nitrogen (50 ms, 9 p.s.i.) using a picospritzer (General Valve) (Fig. 2d). Spontaneous EPSCs were recorded in the presence of 100 μ M picrotoxin and 5 μ M SR-95531. Averages of mEPSCs were collected from at least 150 events.

Outside-out patches were removed from cell bodies of OPCs and solutions were rapidly applied, as described²². Patch currents were filtered at 5 kHz and sampled at 50 kHz, and were made with the Cs-methanesulphonate-based internal solution. The external solution contained (in mM): 135 NaCl, 2.5 KCl, 2.5 CaCl₂, 1.3 MgCl₂, 20 HEPES; pH 7.2.

All experiments were performed at 22–24 °C. Data are expressed as mean $\pm$ standard error, except where noted. Coefficient of variation was calculated as: (standard deviation/mean) $\times 100$.

Histology

OPCs were filled with biocytin (0.15%). The seal was then destroyed with large negative current pulses (to prevent removal of the cell body with the pipette), and the slice fixed. For electron microscopy, slices were fixed in 3.4% paraformaldehyde, 1.25% glutaraldehyde in

0.1 M phosphate buffer (pH 7.4), freeze-thawed and resectioned at 60 μm . Biocytin was visualized with an avidin-biotinylated-HRP using diaminobenzidine or 1 nm gold-conjugated streptavidin (Nanoprobes). Electron microscopic sections were contrasted with lead citrate. One physiologically identified OPC (P21 rat) was fixed 4 h after slice preparation and evaluated in 23 continuous electron microscopy serial sections, measuring the circumference of its processes. The sections, taken at a level of about half radius of the volume occupied by the cell, covered a depth of 1.61 μm and contained 101 μm^2 of OPC membrane, estimated by multiplying the circumference of processes by the estimated section thickness of 70 nm. Assuming isotropic distribution of the processes (total diameter, 65 μm), about 2.5% of the cell was sampled. Distance from axo-spinous synapses to OPC processes was measured in a straight line for distances less than 150 nm without any intervening cellular elements.

For immunocytochemistry, slices were fixed in 4% paraformaldehyde, 0.05% glutaraldehyde in 0.1 M phosphate buffer (pH 7.4), and resectioned at 60 μm . Antibodies to NG2 were provided by W. B. Stallcup and A. Nishiyama. Anti-rat-NG2 antibodies (ab-NG2EC, 2.8 $\mu\text{g ml}^{-1}$) were raised in rabbit to a polypeptide representing the extracellular domain (residues 1–2,223)²⁷. Monoclonal antibodies (D4, culture-supernatant, diluted 1:50) were raised to rat NG2²⁸. Both antibodies label a broad band of relative molecular mass (M_r) of ~400K to 800K on western blots of extracts from rat brain and also from transfected cell lines expressing rat NG2 (A. Nishiyama, personal communication). Guinea pig antibodies to GFAP (1:500) were obtained from Advanced Immunochemicals. Triple fluorescence immunolabelling was performed²⁹ using AMCA-conjugated streptavidin (Vector Labs) for biocytin, Alexa 488-conjugated anti-rabbit or anti-mouse IgG (Molecular Probes) for NG2, and CY3-conjugated anti-guinea pig IgG (Jackson Labs) for GFAP.

Two adult (>P60) and two developing (P19, 21) male rats (Wistar) were anaesthetized, fixed and immunoreacted with ab-NG2EC (1:500) using silver intensified pre-embedding immunogold reaction (1.4 nm gold conjugated secondary antibody, Nanoprobes) as described³⁰. Because these antibodies recognize an extracellular epitope, only the silver intensified immunogold labelling method is suitable for identifying glial processes in the neuropil. No labelling was detected when specific antibodies were replaced with normal serum of the same species or when primary antibody was omitted.

Materials

Pardaxin and α -latrotoxin were purchased from Alamone Labs. All other drugs were purchased from RBI (USA).

Received 17 November 1999; accepted 23 February 2000.

- Steinhauser, C. & Gallo, V. News on glutamate receptors in glial cells. *Trends Neurosci.* **19**, 339–345 (1996).
- Miller, R. H. Oligodendrocyte origins. *Trends Neurosci.* **19**, 92–96 (1996).
- Palay, S. L. & Chan-Palay, V. *Cerebellar Cortex, Cytology and Organization* (Springer, New York, 1974).
- Raff, M. C., Miller, R. H. & Noble, M. A glial progenitor cell that develops in vitro into an astrocyte or an oligodendrocyte depending on culture medium. *Nature* **303**, 390–396 (1983).
- Patneau, D. K., Wright, P. W., Winters, C., Mayer, M. L. & Gallo, V. Glial cells of the oligodendrocyte lineage express both kainate- and AMPA-preferring subtypes of glutamate receptor. *Neuron* **12**, 357–371 (1994).
- Wyllie, D. J., Mathie, A., Symonds, C. J. & Cull-Candy, S. G. Activation of glutamate receptors and glutamate uptake in identified macroglial cells in rat cerebellar cultures. *J. Physiol. (Lond.)* **432**, 235–258 (1991).
- Barres, B. A. & Raff, M. C. Control of oligodendrocyte number in the developing rat optic nerve. *Neuron* **12**, 935–942 (1994).
- Gallo, V. *et al.* Oligodendrocyte progenitor cell proliferation and lineage progression are regulated by glutamate receptor-mediated K^+ channel block. *J. Neurosci.* **16**, 2659–2670 (1996).
- McDonald, J. W., Levine, J. M. & Qu, Y. Multiple classes of the oligodendrocyte lineage are highly vulnerable to excitotoxicity. *Neuroreport* **9**, 2757–2762 (1998).
- Bergles, D. E., Diamond, J. S. & Jahr, C. E. Clearance of glutamate inside the synapse and beyond. *Curr. Opin. Neurobiol.* **9**, 293–298 (1999).
- Krieger, S. & Chiu, S. Y. Calcium signaling of glial cells along mammalian axons. *J. Neurosci.* **13**, 4229–4245 (1993).
- Ong, W. Y. & Levine, J. M. A light and electron microscopic study of NG2 chondroitin sulfate proteoglycan-positive oligodendrocyte precursor cells in the normal and kainate-lesioned rat hippocampus. *Neurosci.* **92**, 83–95 (1999).
- Dingledine, R., Borges, K., Bowie, D. & Traynelis, S. F. The glutamate receptor ion channels. *Pharmacol. Rev.* **51**, 7–61 (1999).
- Zerangue, N. & Kavanaugh, M. P. Flux coupling in a neuronal glutamate transporter. *Nature* **383**, 634–637 (1996).
- Dunwiddie, T. V. The physiological role of adenosine in the central nervous system. *Int. Rev. Neurobiol.* **27**, 63–139 (1985).
- Amaral, D. G. & Witter, M. P. The three dimensional organization of the hippocampal formation: a review of anatomical data. *Neurosci.* **31**, 571–591 (1989).
- Lisman, J. E. & Harris, K. M. Quantal analysis and synaptic anatomy - integrating two views of hippocampal plasticity. *Trends Neurosci.* **16**, 141–147 (1993).
- Renner, P., Caratsch, C. G., Waser, P. G., Lazarovici, P. & Primor, N. Presynaptic effects of the pardaxins, polypeptides isolated from the gland secretion of the flatfish *Pardachirus Marmoratus*. *Neurosci.* **23**, 319–325 (1987).
- Hessler, N. A., Shirke, A. M. & Malinow, R. The probability of transmitter release at a mammalian central synapse. *Nature* **366**, 569–572 (1993).
- Dzubay, J. A. & Jahr, C. E. The concentration of synaptically released glutamate outside of the climbing fiber-purkinje cell synaptic cleft. *J. Neurosci.* **19**, 5265–5274 (1999).
- Ventura, R. & Harris, K. M. Three-dimensional relationships between hippocampal synapses and astrocytes. *J. Neurosci.* **19**, 6897–6906 (1999).

- Bergles, D. E., Dzubay, J. A. & Jahr, C. E. Glutamate transporter currents in bergmann glial cells follow the time course of extrasynaptic glutamate. *Proc. Natl Acad. Sci. USA* **94**, 14821–14825 (1997).
- Wang, S. *et al.* Notch receptor activation inhibits oligodendrocyte differentiation. *Neuron* **21**, 63–75 (1998).
- Keirstead, H. S., Levine, J. M. & Blakemore, W. F. Response of the oligodendrocyte progenitor cell population (defined by NG2 labelling) to demyelination of the adult spinal cord. *Glia* **22**, 161–170 (1998).
- Wolszijk, G. Chronic stage multiple sclerosis lesions contain a relatively quiescent population of oligodendrocyte precursor cells. *J. Neurosci.* **18**, 601–609 (1998).
- Bergles, D. E. & Jahr, C. E. Synaptic activation of glutamate transporters in hippocampal astrocytes. *Neuron* **19**, 1297–1308 (1997).
- Tillet, E., Ruggiero, F., Nishiyama, A. & Stallcup, W. B. The membrane-spanning proteoglycan NG2 binds to collagens V and VI through the central nonglobular domain of its core protein. *J. Biol. Chem.* **272**, 10769–10776 (1997).
- Stallcup, W. B., Dahlin, K. & Healy, P. Interaction of the NG2 chondroitin sulfate proteoglycan with type VI collagen. *J. Cell Biol.* **111**, 3177–3188 (1990).
- Reyes, A. *et al.* Target-cell-specific facilitation and depression in neocortical circuits. *Nature Neurosci.* **1**, 279–285 (1998).
- Baude, A. *et al.* The metabotropic glutamate receptor (mGluR1 α) is concentrated at perisynaptic membrane of neuronal subpopulations as detected by immunogold reaction. *Neuron* **11**, 771–787 (1993).

Acknowledgements

We thank P. Cobden and P. Jays for assistance. This work was supported by the UK Medical Research Council (P.S.) and the NIH (C.E.J.).

Correspondence and requests for material should be addressed to D.E.B. (e-mail: berglesd@ohsu.edu).

Fringe forms a complex with Notch

Bong-Gun Ju*†, Sangyun Jeong*†, Eunkyung Bae*†‡, Seogang Hyun*§, Sean B. Carroll||, Jeongbin Yim* & Jaeseob Kim*§

* National Creative Research Initiative Center for Genetic Reprogramming, Institute of Molecular Biology and Genetics, Seoul National University, Shinlim-Dong, Kwanak-Ku, Seoul 151-742, Korea

|| Howard Hughes Medical Institute and Laboratory of Molecular Biology, University of Wisconsin at Madison, 1525 Linden Drive, Madison, Wisconsin 53706, USA

§ Department of Biological Sciences, Korea Advanced Institute of Science & Technology, Kusong-dong, Yusong-ku, Taejeon, 305-701, Korea

† These authors contributed equally to this work

The Fringe protein of *Drosophila* and its vertebrate homologues function in boundary determination during pattern formation^{1–9}. Fringe has been proposed to inhibit Serrate–Notch signalling but to potentiate Delta–Notch signalling¹⁰. Here we show that Fringe and Notch form a complex through both the Lin–Notch repeats and the epidermal growth factor repeats 22–36 (EGF22–36) of Notch when they are co-expressed. The *Abruptex*^{59b} (*Ax*^{59b}) and *Ax*^{M1} mutations, which are caused by missense mutations in EGF repeats 24 and 25, respectively, abolish the Fringe–Notch interaction through EGF22–36, whereas the *l(1)N^B* mutation in the third Lin–Notch repeat of Notch abolishes the interaction through Lin–Notch repeats. *Ax* mutations also greatly affect the Notch response to ectopic Fringe *in vivo*. Results from *in vitro* protein mixing experiments and subcellular colocalization experiments indicate that the Fringe–Notch complex may form before their secretion. These findings explain how Fringe acts cell-autonomously to modulate the ligand preference of Notch and why the Fringe–Notch relationship is conserved between phyla and in the development of very diverse structures.

Fringe (Fng), an extracellular protein, determines the boundary of two different cell populations during the development of diverse

‡ Present address: Department of Ophthalmology & Visual Science, Catholic Research Institute of Medical Science, The Catholic University of Korea, Seocho-Gu, Seoul 137-701, Korea.

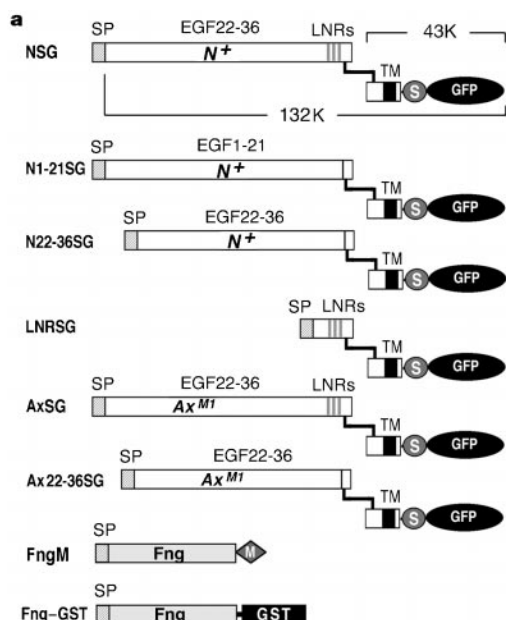


Figure 1 The Fringe and Notch derivatives used in this study. **a**, The Fng and N derivative proteins expressed in transgenic flies using Gal4/UAS system²⁹ with *hsp70-Gal4* driver or in transfected cell lines under the metallothioneine promoter. Note that the processing of NSG in *trans*-Golgi produces the *M*_r 89K and 43K polypeptides which are linked by disulphide bond(s). Constructs: NSG, EGF repeat 22 to the transmembrane domain (TM) of Notch fused to S tag (S) and GFP; N1-21SG, EGF 1-21 of Notch fused to S tag and GFP; N22-36SG, EGF 22-86 of Notch fused to S tag and GFP; LNRSG, LNRs of Notch fused to S tag and GFP; AxSG, Ax mutant form of NSG; Ax22-36SG, Ax mutant form of N22-36SG; FngM, Fng tagged with Myc epitope (M); Fng-GST, Fng fused to GST. All constructs contain a signal peptide sequence (SP). **b**, Fluorescence microscopic images of stably transfected cell lines expressing NSG and its derivatives.

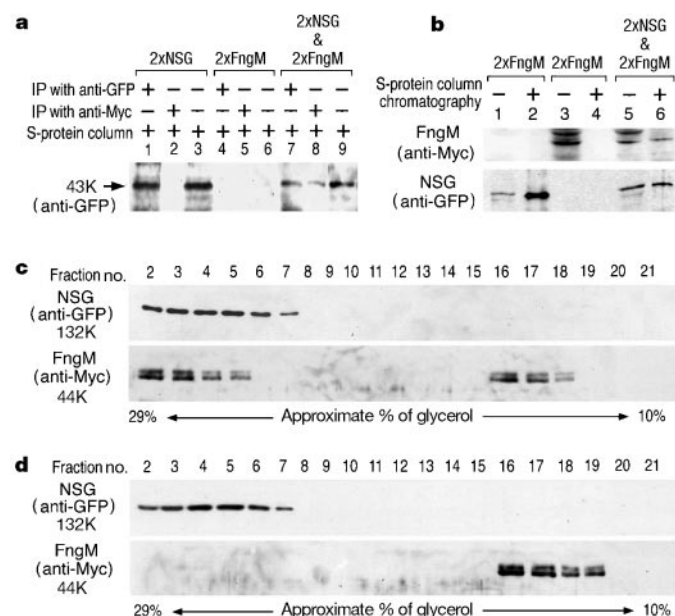
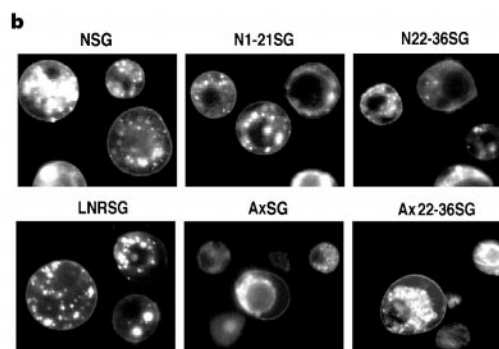


Figure 2 Fng and Notch are present as a complex. Fng and Notch derivatives were expressed in *Drosophila* larvae (**a**, **b**) or in S2 cells (**c**–**g**). Notch derivatives, FngM and Fng-GST were detected by western blot with anti-GFP, anti-Myc or anti-GST antibody, respectively. 2x, two copies of transgenes; &, co-expression; x, *in vitro* mixing of the Fng and Notch derivatives expressed separately; +/–, with (+) or without (–) the treatment denoted on the left ends of the lines; IP, immunoprecipitation. **a**, Co-immunoprecipitation of NSG with FngM. **b**, Co-purification of FngM with NSG on S-protein column chromatography. Upper panel, western blot with anti-Myc antibody; lower panel, with anti-GFP antibody. **c**, FngM co-migrates with NSG in glycerol-gradient centrifugation. Upper panel, western blot with anti-GFP antibody; lower panel, with anti-Myc antibody. **d**, When NSG and FngM are expressed separately and mixed *in vitro*, FngM does not co-

migrate with NSG in glycerol-gradient centrifugation. **e**, Co-purification of Notch derivatives with Fng-GST on glutathione column chromatography. When co-expressed in the same cells, NSG, N22-36SG, LNRSG and AxSG are co-purified with Fng-GST on glutathione column chromatography (lanes 12, 16, 18, 22, respectively) whereas N1-21SG (lane 14) and Ax22-36SG (lane 26) are not. The affinity of Fng for N22-36SG appears to be lower than that for LNRSG. None of the Notch derivatives are co-purified with Fng-GST when they are expressed in separate cells and mixed *in vitro* (lanes 27–34). **f**, The A59b22-36SG (A59b22-36SG) and the I(1)^N form of LNRSG (NB-LNRSG). **g**, A59b22-36SG and NB-LNRSG are not co-purified with Fng-GST. Extracts from approximately 2×10^7 cells were used for each lane in **e** and **g**.

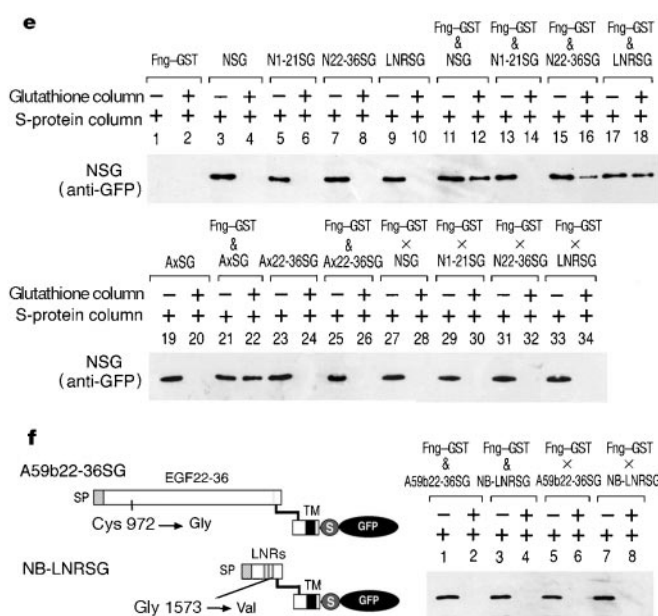


Figure 2e Co-purification of Notch derivatives with Fng-GST on glutathione column chromatography. When co-expressed in the same cells, NSG, N22-36SG, LNRSG and AxSG are co-purified with Fng-GST on glutathione column chromatography (lanes 12, 16, 18, 22, respectively) whereas N1-21SG (lane 14) and Ax22-36SG (lane 26) are not. The affinity of Fng for N22-36SG appears to be lower than that for LNRSG. None of the Notch derivatives are co-purified with Fng-GST when they are expressed in separate cells and mixed *in vitro* (lanes 27–34). **f**, The A59b22-36SG (A59b22-36SG) and the I(1)^N form of LNRSG (NB-LNRSG). **g**, A59b22-36SG and NB-LNRSG are not co-purified with Fng-GST. Extracts from approximately 2×10^7 cells were used for each lane in **e** and **g**.

structures not only in *Drosophila* but also in vertebrates^{1–10}. The identification of the proteins that physically interact with Fng is important to understand the molecular mechanisms of Fng function. As most known Fng-mediated developmental processes require Notch signalling^{2–7,10–15}, Notch is a strong candidate for Fng-interacting proteins.

To test whether Fng binds Notch, we performed a series of biochemical experiments focusing on the EGF-like repeats 22 to 36 and Lin–Notch repeats (LNRs) of Notch, to which most Ax mutations and antineurogenic mutations map^{16–19}. A chimaeric Notch, named NSG, which is EGF repeat 22 to the transmembrane domain of Notch fused to an S tag and green fluorescent protein (GFP), and Fng tagged with the Myc epitope (FngM) were expressed separately or together in transgenic *Drosophila* (Fig. 1a). The extracts of the transgenic larvae were immunoprecipitated with anti-GFP or anti-Myc antibody and then purified by S-protein affinity column chromatography (Fig. 2a). This ensured that all of the positive bands shown in subsequent western blots with anti-GFP antibody contained an S tag as well as a GFP tag. Because Notch is processed into two polypeptides linked by disulphide bonds in the *trans*-Golgi^{20,21}, the NSG protein should yield two polypeptides by SDS–polyacrylamide gel electrophoresis (see Fig. 1a). The first product is an amino-terminal peptide of relative molecular mass

89,000 (M_r 89K) which is untagged. The second polypeptide is a carboxy-terminal peptide (NSG-C) of M_r 43K which contains an S tag and GFP. As shown in Fig. 2a, the eluents of the S-protein affinity column from extracts of animals expressing 2xNSG or 2xNSG plus 2xFngM all yield a positive 43K band as expected (Fig. 2a, lanes 1, 3, 7, 9). When the immunoprecipitations were carried out with anti-Myc antibody, which specifically bound to FngM, the extract from the animal co-expressing NSG and FngM yielded the co-precipitated 43K NSG-C peptide (Fig. 2a, lane 8).

The Fng–Notch interaction was confirmed by co-purification of the FngM protein with the NSG protein by S-protein affinity chromatography (Fig. 2b), as well as by glycerol-gradient centrifugation analysis (Fig. 2c). As previously reported²², Fng is present in two forms (Fig. 2b upper panel, lanes 3, 5). When FngM was co-expressed with NSG, the higher mobility form of FngM (Fig. 2b upper panel, lower band of lane 5) was well retained with NSG on the S-protein column (Fig. 2b upper panel, lane 6).

In glycerol-gradient centrifugation analysis, both the high and low mobility forms of FngM were found in fractions 2–5 where NSG was present (Fig. 2c, lower panel). Because unmodified monomeric FngM has M_r 44K, and monomeric NSG has M_r 132K, this colocalization of the high and low mobility forms of FngM with NSG in the glycerol-gradient implies that they are present as a complex. The free FngM that is not bound to NSG was detected in fractions 16–18.

To map the domains of Notch that interact with Fng, we

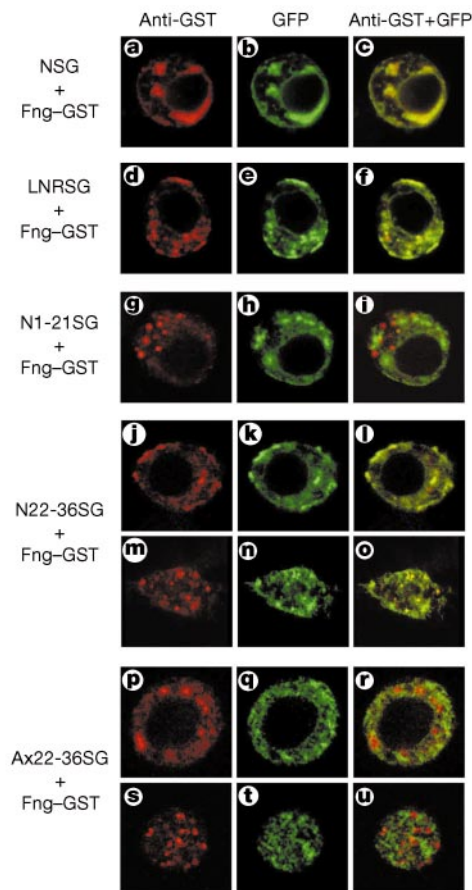


Figure 3 Colocalization of Fng and Notch protein derivatives in transfected cells. S2 cells expressing Fng–GST and Notch derivatives indicated in the left of the panels were fixed and stained with anti-GST antibody. Fringe was detected with LRSC-coupled anti-mouse IgG antibody (red). The expressions of Notch derivatives were monitored with GFP expression (green). The panels at the far right are the merged images of the left two panels. **m–o, s–u**, The images scanned along the focal planes near the surfaces of the cells shown in **j–i** and **p–r**, respectively. To minimize the false signal in inappropriate channels, we used a multi-tracking method with an LSM510 confocal microscope (Carl Zeiss).

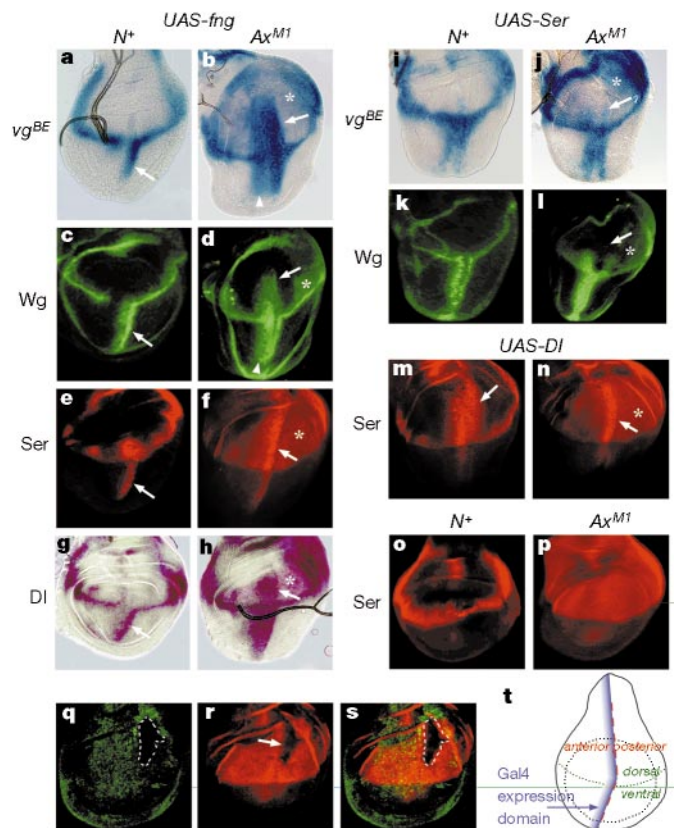


Figure 4 Ax mutations affect the Fng–Notch interaction *in vivo*. **a–p**, The expression of *vgBE-lacZ* (blue), *Wg* (green), *Ser* (red) and *DI-lacZ* (purple) are shown in mid-third instar wing discs of wild-type (**a, c, e, g, i, k, m, o**) and *AxM1* (**b, d, f, h, j, l, n, p**) without (**o, p**) or with transgenes expressing Fng (**a–h**), *Ser* (**i–l**), and *DI* (**m, n**) in the *ptc* pattern. **q–r**, *fng*¹³ homozygous null clones, which are marked by the absence of Myc staining (**q**, green), do not express *Ser* (**r**, red) in an *AxM1* mutant disc. The *fng*¹³ clone is outlined. **s**, Merged image of **q** and **r**. **t**, The ectopic expression domain of transgenes by UAS/ptc–Gal4 technique²⁹.

expressed various forms of Notch derivatives with an S tag and GFP (Fig. 1) in stably transfected S2 cell lines with or without Fng–GST (Fng fused to glutathione S-transferase). Ax^{M1} is a missense mutation of Cys 999 to tyrosine (S.H. and J.K., unpublished data; ref. 18). The extracts from these cell lines were loaded onto glutathione-Sepharose column. If NSG, or its derivatives, formed a complex with Fng, it should be retained on the column with Fng–GST. NSG, LNRSG, N22–36SG and AxSG were retained with the Fng–GST protein on glutathione-Sepharose (Fig. 2e, lanes 12, 16, 18, 22), but N1–21SG and Ax22–36SG were not (Fig. 2e, lanes 14, 26). These data suggest that Fng specifically binds to Notch through the LNRs as well as the EGF22–36 and that Ax^{M1} mutation abolishes the Fng interaction with EGF22–36 of Notch but does not affect the Fng interaction with the LNRs.

To test further the specificity of the Fng–Notch interaction, we expressed two additional mutant forms of N derivatives in S2 cells; the Ax^{59b} mutant form of N22–36SG (A59b22–36SG) and the $l(1)N^B$ mutant form of LNRSG (NB–LNRSG) (Fig. 2f). The Ax^{59b} mutation is a Cys 972 to glycine substitution in EGF repeat 24; the $l(1)N^B$ mutation is a Gly 1572 to valine substitution in the third LNR^{16,19}. As expected, both mutations abolish the interaction of Fng with the EGF22–36 and with the LNRs of Notch, respectively (Fig. 2g). These data suggest that the Fng–Notch interaction is specific.

The subcellular colocalization of Fng–GST with the Notch derivatives also support the specific Fng–Notch interaction. We examined the subcellular localization of Fng–GST and the Notch derivatives which were co-expressed in S2 cells. The subcellular distribution of Fng mostly overlapped with that of the Notch derivatives which interacted with Fng in biochemical experiments (NSG, N22–36SG and LNRSG; Fig. 3a–f and Fig. 3j–o). In contrast,

N1–21SG and Ax22–36SG are mainly not colocalized with Fng and did not interact (Fig. 3g–i, p–u). These data indicate that Fng may be present as a complex with Notch even before its secretion.

Consistent with these observations, when Fng–GST and the Notch derivatives were expressed separately and mixed *in vitro*, none of the Notch derivatives co-purified with Fng–GST on a glutathione-Sepharose column (Fig. 2e, lanes 27–34). In addition, when FngM and NSG were expressed separately and mixed *in vitro* and then analysed by glycerol-gradient centrifugation, FngM was mainly found in fractions 16–19 (Fig. 2d, lower panel), whereas NSG was found in fractions 2–7 (Fig. 2d, upper panel). This implies that Fng and Notch do not form a complex if they are not expressed in the same cells simultaneously. These data are consistent with the finding that Fng is not detected on cells expressing Notch when Fng protein is added exogenously¹⁰. Therefore, Fng–Notch complex formation may occur before secretion, probably within the secretory pathway. This explains why the secretory Fng protein acts cell-autonomously in Notch-expressing cells. It is unlikely to modify Notch found on the surface of neighbouring cells or cells that may be reached by secreted Fng. Fng has been proposed as a glycosyltransferase on the basis of sequence similarity²³. Fng may bind Notch and may modify the glycosylation of Notch during its secretion in the endoplasmic reticulum (ER) or Golgi. A comparison of the glycosylation status between the Fng–Notch complex and Notch alone will be of considerable interest.

Genetic evidence from Ax mutants suggests that the Notch EGF 22–36 are involved in the Notch–Fng interaction *in vivo*. In wild-type mid-third *Drosophila* wing imaginal discs, the expression of Serrate (Ser), the *vestigial* (*vg*) boundary enhancer-lacZ (*vgBE*) and *wingless* (*wg*) are upregulated by Fng and Notch near or at the dorsoventral (D–V) boundary but not in dorsal interior cells². In contrast, in Ax^{M1} mutants Ser and *vgBE* expression are expanded into all dorsal cells, including interior cells (Fig. 4p, and asterisks in Fig. 4b, d; ref. 24). This expansion may be due to disturbance of the Fng–Notch interaction through the EGF22–36 of Notch by the Ax mutation.

To test this, we ectopically expressed Fng in Ax^{M1} mutant wing discs using *UAS-fng* and a *patched* (*ptc*)–Gal4 driver (Fig. 4t). In wild-type imaginal discs, ectopic expression of Fng in the *ptc* pattern induces *vgBE* (Fig. 4a), *Wg* (Fig. 4c), Ser (Fig. 4e) and Delta (DI) (Fig. 4g) only in ventral cells in a single stripe along the A/P boundary (arrows)^{2,10}. In Ax^{M1} mutants, however, Ser, *Wg*, *vgBE* and DI expression were strongly induced by ectopic Fng even in dorsal cells (Fig. 4b, d, f, h, arrows). The level of Ser expression in the dorsal cells expressing ectopic Fng is much higher than that in other dorsal cells (compare the areas indicated with arrow and asterisk in Fig. 4f). The effects of ectopic Fng expression on Ser, *Wg*, *vgBE*, and DI expression are also similar in Ax^{28} homozygous and Ax^{16172}/Ax^{28} heterozygous mutant wing discs although the extent of the effect varies (data not shown). Ax mutations do not affect the response of dorsal or ventral cells to Ser or DI as significantly as their response to Fng (Fig. 4i–n). This implies that the induction of Ser, DI and *vgBE* by Fng in Ax mutant dorsal cells may not be mediated by Ser.

To confirm the effects of Fng on Ser expression in Ax^{M1} mutants, we generated homozygous *fng* null mitotic clones in the Ax^{M1} mutant background. Notably, homozygous *fng*[−] dorsal cells of Ax^{M1} mutant wing discs do not express Ser (Fig. 4q–s, arrows). Ser expression in the interior dorsal cells of Ax^{M1} mutant therefore requires Fng function, and the Ax^{M1} mutation disrupts the normal downstream regulatory effects of the Fng–Notch interaction.

The LNRs of Notch function as a repressor domain in Notch signalling^{19,25,26}. As Ax mutations prevent EGF22–36 of Notch from binding to Fng but do not affect Fng binding to the LNRs of Notch (Fig. 2e, lanes 19–26), one of the possible mechanisms of Fng-dependent Notch target gene activation in Ax mutants is that the Ax mutation allows Fng to bind to LNRs alone rather than to EGF 22–36. This may dampen or silence the repressor function of the LNRs,

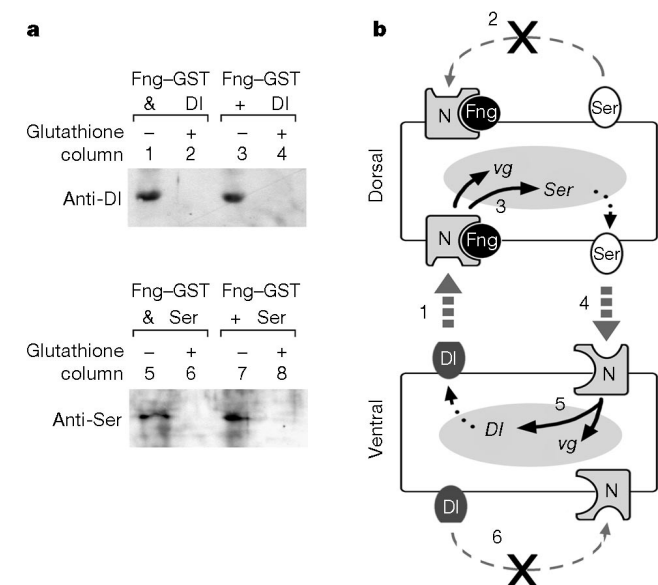


Figure 5 Modulation of Notch signalling by Fng. **a**, Fng does not interact with DI or Ser. DI or Ser which is co-expressed with Fng–GST (denoted with '&') or expressed in S2 cells separately and mixed with Fng–GST *in vitro* ('+') were detected by western blot before (odd numbered lanes) or after (even numbered lanes) purification on glutathione-Sepharose column. Neither DI nor Ser co-purified with Fng–GST on the column. **b**, Model of Fng–Notch signalling in *Drosophila* wing imaginal discs. Notch is secreted to the membrane as a complex with Fng in the dorsal cells but not in the ventral cells. Notch in complex with Fng can respond to DI (1) but not to Ser (2). The Fng–Notch complex activated by DI induces Ser, *vg* and *wg* in the dorsal cells (3). The Ser induced by Fng–Notch complex in the dorsal cells then specifically activates Notch in the ventral boundary cells (4). The active Notch signal in the ventral cells turns on *DI* as well as, for example, *vg* and *wg* (5). The Notch in the ventral boundary cells that is not complexed with Fng cannot be activated by DI (6).

and might, in turn, make Notch more sensitive to D Δ . Alternatively, Ax mutations may mimic the conformation of Notch bound with a ligand and Fng may help this form of Notch to be converted to a fully activated form by antagonizing the repressor function of the LNRs.

Fringe has been proposed to execute its boundary determining function by inhibiting the Notch response to Ser and potentiating the Notch response to D Δ (ref. 10). Because Fng does not bind Ser or D Δ (Fig. 5a), modulation of Notch signalling by Fng is directly mediated by the complex formation of Fng and Notch during their secretory transits. Notch bound to Fng may have preferential affinity or sensitivity to Delta, whereas free Notch may have a higher affinity or sensitivity to Ser (Fig. 5b). Upon D Δ binding to the Fng–Notch complex, Fng may also antagonize the repressor function of the LNRs and facilitate the activation of Notch signalling. □

Methods

The protein expression constructs for S2 cells were made using pMthY vector¹⁰. Culture of S2 cells, transfection with dimethyldioctadecylammonium bromide (DDAB; Sigma), induction with CuSO₄ and selection of stably transfected cells with hygromycin B were done as described^{27,28}. For extract preparation, the samples homogenized in 20 mM Tris, pH 7.6, 150 mM NaCl, 1 mM EGTA, 1 mM DTT, 1% Triton X-100 were centrifuged and the supernatant was ultrafiltered to replace the buffer containing 0.2% Triton X-100 using Centricon-30 (Amicon). Ultracentrifugation for glycerol-gradient analysis was done at 45,000g for 42 h at 4 °C. The *fng* null clones were generated by heat shock for 1 h at 37 °C to the progenies of AxM1/Fm6; *Fng*¹³-FRT/TM6b females crossed with Fm6;FLPase³⁸/Bc;Gla;80_M males.

Received 24 January; accepted 7 March 2000.

- Irvine, K. & Wieschaus, E. *fringe*, a boundary-specific signalling molecule, mediates interactions between dorsal and ventral cells during *Drosophila* wing development. *Cell* **79**, 595–606 (1994).
- Kim, J., Irvine, K. D. & Carroll, S. B. Cell recognition, signal induction, and symmetrical gene activation at the dorsal-ventral boundary of the developing *Drosophila* wing. *Cell* **82**, 795–802 (1995).
- Papayannopoulos, V., Tomlinson, A., Panin, V. M., Rauskolb, C. & Irvine, K. D. Dorsal–ventral signaling in the *Drosophila* eye. *Science* **281**, 2031–2034 (1998).
- Cho, K. O. & Choi, K. W. Fringe is essential for mirror symmetry and morphogenesis in the *Drosophila* eye. *Nature* **396**, 272–276 (1998).
- Dominguez, M. & de Celis, J. F. A dorsal/ventral boundary established by Notch controls growth and polarity in the *Drosophila* eye. *Nature* **396**, 276–278 (1998).
- Esteban, C. R. *et al.* Radical fringe position the apical ectodermal ridge at the dorsoventral boundary of the vertebrate limb. *Nature* **386**, 360–365 (1998).
- Laufer, E. D. *et al.* Expression of Radical fringe in limb-bud ectoderm regulates apical ectodermal ridge formation. *Nature* **386**, 366–373 (1997).
- Zhang, N. & Gridley, T. Defects in somite formation in lunatic fringe-deficient mice. *Nature* **394**, 374–377 (1998).
- Evrard, Y. A., Lun, Y., Aulehla, A., Gan, L. & Johnson, R. L. Lunatic fringe is an essential mediator of somite segmentation and patterning. *Nature* **394**, 377–381 (1998).
- Panin, V. M., Papayannopoulos, V., Wilson, R. & Irvine, K. D. Fringe modulates Notch–ligand interactions. *Nature* **387**, 908–912 (1997).
- Kim, J. *et al.* Integration of positional signals and regulation of wing formation and identity by *Drosophila* vestigial gene. *Nature* **382**, 133–138 (1996).
- Kim, J., Magee, J. & Carroll, S. B. Intercompartmental signaling and the regulation of vestigial expression at the dorsoventral boundary of the developing *Drosophila* wing. *Cold Spring Harb. Symp. quant. Biol.* **62**, 283–291 (1997).
- Doherty, D. *et al.* Delta is an ventral to dorsal signal complementary to Serrate, another Notch ligand, in *Drosophila* wing formation. *Genes Dev.* **10**, 421–434 (1996).
- de Celis, J. F. & Bray, S. Feed-back mechanisms affecting Notch activation at the dorsoventral boundary in the *Drosophila* wing. *Development* **124**, 3241–3251 (1997).
- Klein, T. & Arias, A. M. Interactions among Delta, Serrate and Fringe modulate Notch activity during *Drosophila* wing development. *Development* **125**, 2951–2962 (1998).
- Kelley, M. R., Kidd, S., Deutsch, W. A. & Young, M. W. Mutations altering the structure of epidermal growth factor-like coding sequences at the *Drosophila* Notch locus. *Cell* **51**, 539–548 (1987).
- de Celis, J. F., Barrio, R., Arco, A. D. & Bellido, A. G. Genetic and molecular characterization of a Notch mutation in its Delta- and Serrate-binding domain in *Drosophila*. *Proc. Natl Acad. Sci. USA* **90**, 4037–4041 (1993).
- Brennan, K., Tateson, R., Lewis, K. & Arias, A. M. A functional analysis of Notch mutations in *Drosophila*. *Genetics* **147**, 177–188 (1997).
- Lyman, D. & Young, M. W. Further evidence for function of the *Drosophila* Notch protein as a transmembrane receptor. *Proc. Natl Acad. Sci. USA* **90**, 10395–10399 (1993).
- Blaumüller, C. M., Qi, H., Zagouras, P. & Artavanis-Tsakonas, S. Intracellular cleavage of Notch leads to a heterodimeric receptor on the plasma membrane. *Cell* **90**, 281–291 (1997).
- Loeate, F. *et al.* The Notch1 receptor is cleaved constitutively by a furin-like convertase. *Proc. Natl Acad. Sci. USA* **95**, 8108–8112 (1998).
- Johnston, S. H. *et al.* A family of mammalian Fringe genes implicated in boundary determination and the Notch pathway. *Development* **124**, 2245–2254 (1997).
- Yuan, Y. P., Schultz, J., Mlodzik, M. & Bork, P. Secreted Fringe-like signaling molecules may be glycosyltransferases. *Cell* **88**, 9–11 (1997).
- de Celis, J. F., Garcia-Bellido, A. & Bray, S. J. Activation and function of Notch at the dorsal–ventral boundary of the wing imaginal disc. *Development* **122**, 359–369 (1996).

- Lieber, T. *et al.* Antineurogenic phenotypes induced by truncated Notch proteins indicate a role in signal transduction and may point to a novel function for Notch in nuclei. *Genes Dev.* **7**, 1949–1965 (1993).
- Kopan, R., Schroeter, E. H., Weintraub, H. & Nye, J. S. Signal transduction by activated mNotch: Importance of proteolytic processing and its regulation by the extracellular domain. *Proc. Natl Acad. Sci. USA* **93**, 1683–1688 (1996).
- Han, K. An efficient DDAB-mediated transfection of *Drosophila* S2 cells. *Nucleic Acids Res.* **24**, 4362–4363 (1996).
- Johansen, H. *et al.* Regulated expression at high copy number allows production of a growth-inhibitory oncogene product in *Drosophila* Schneider cells. *Genes Dev.* **3**, 882–889 (1989).
- Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* **118**, 401–415 (1993).

Acknowledgements

We would like to thank G. Panganiban for helpful comments; D. Kim, A. Hudson and J. Wilson for help in manuscript preparation; P. Fernandez and A. Garcia-Bellido for Ax mutant fly stocks; T. Klein, H. Marc, and A. Martinez-Arias for UAS-D Δ and D Δ -lacZ fly stocks; K. Basler for the FLPase38 stock; E. Knust for Ser antibody; R. Fleming for the *hsp70-Gal4* stock and the construct to express Ser in S2 cells; Muskavitch for the construct to express D Δ in S2 cells. S.B. is an investigator of the Howard Hughes Medical Institute. This work was supported by KAIST BK21 program to J. Kim and a grant to J. Yim from Creative Research Initiatives of the Korean Ministry of Science and Technology.

Correspondence and requests for materials should be addressed to Jaeseob Kim (e-mail: kjaeseob@sorak.kaist.ac.kr).

ATF-2 has intrinsic histone acetyltransferase activity which is modulated by phosphorylation

Hiroaki Kawasaki[†], Lou Schiltz[‡], Robert Chiu[§], Keiichi Itakura^{||}, Kazunari Taira[¶], Yoshihiro Nakatani^{‡#} & Kazunari K. Yokoyama^{*}

^{*} Tsukuba Life Science Center, The Institute of Physical and Chemical Research (RIKEN), Tsukuba 305-0074, Japan

[†] National Institute for Advanced Interdisciplinary Research and National Institute of Bioscience and Human Technology, Agency of Industrial Science & Technology, MITI, Tsukuba 305-0006, Japan

[‡] Laboratory of Molecular Growth Regulation, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892, USA

[§] School of Medicine/School of Dentistry, Surgical Oncology/Oral Biology and Medicine, University of California, Los Angeles, California 90095-1668, USA

^{||} Department of Molecular Genetics, Beckman Research Institute of the City of Hope, 1450 East Duarte, Duarte, California 91010, USA

[¶] Department of Chemistry and Biotechnology, Graduate School of Engineering, University of Tokyo, Tokyo 113-8656, Japan

Transcription factors carry functional domains, which are often physically distinct, for sequence-specific DNA binding, transcriptional activation and regulatory functions. The transcription factor ATF-2 is a DNA-binding protein that binds to cyclic AMP-response elements (CREs), forms a homodimer or heterodimer with c-Jun, and stimulates CRE-dependent transcription^{1–3}. Here we report that ATF-2 is a histone acetyltransferase (HAT), which specifically acetylates histones H2B and H4 *in vitro*. Motif A, which is located in the HAT domain, is responsible for the stimulation of CRE-dependent transcription; moreover, in response to ultraviolet irradiation, phosphorylation of ATF-2 is accompanied by enhanced HAT activity of ATF-2 and CRE-dependent transcription. These results indicate that phosphorylation of ATF-2 controls its intrinsic HAT activity and its action on CRE-dependent transcription. ATF-2 may represent a new class of sequence-specific factors, which are able to activate transcription by direct effects on chromatin components.

Present address: Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA.

Various coactivators of transcription, such as GCN/5 (ref. 4), p300/CBP (refs 5, 6), PCAF (ref. 7), ACTR (ref. 8), SRC-1 (ref. 9), TAF_{II}250 (ref. 10) and p/CIP (ref. 11), have been identified as histone acetyltransferases (HATs). The ability of the coactivator to catalyse acetylation may be linked to its ability to serve as a transcriptional activator that is involved in the regulation of expression of many genes and in the remodelling of nucleosomes^{12,13}. We previously showed that, upon treatment of F9 cells with retinoic acid or adenovirus E1A (ref. 14), ATF-2 interacts with region p300 of the coactivator, a region that includes the active site of the HAT domain, to control transcription of the *c-jun* gene. We therefore postulated that ATF-2 affects the intrinsic HAT activity of p300 (refs 5, 6). We

monitored the HAT activity of p300 *in vitro* in the presence and absence of ATF-2 (ref. 15). As we reported previously^{5,6}, the HAT domain of p300 is responsible for the HAT activity of p300 *in vitro*, whereas glutathione *S*-transferase (GST) itself has no HAT activity (Fig. 1a). The HAT activity of p300 was enhanced after incubation with ATF-2. Unexpectedly, significant HAT activity was still observed when ATF-2 was incubated with p300 that lacked the HAT domain and thus had no detectable HAT activity *per se* (Fig. 1a). This suggested that the residual HAT activity was due to ATF-2.

We then examined the enhancement of HAT activity by ATF-2 *in vivo*. We assayed the HAT activity of immunoprecipitates obtained

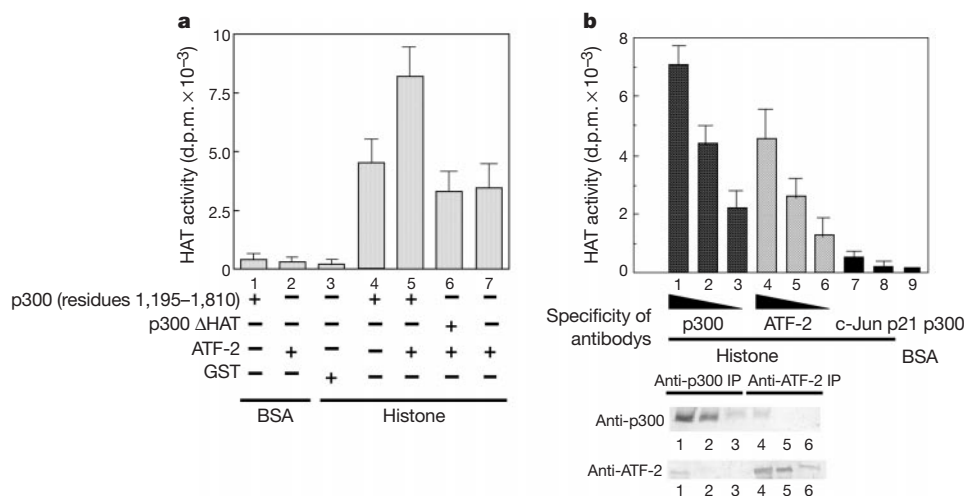


Figure 1 HAT activities of p300 and ATF-2. **a**, GST-fusion proteins (2–20 pmol) corresponding to p300 (residues 1195–1810)⁶, p300ΔHAT (ref. 6), and GST–ATF-2 (ref. 15) were tested for HAT activity. **b**, HAT activity was measured in the indicated immunoprecipitates with antibodies specific for p300, c-Jun, p21 and ATF-2, respectively. Amounts of p300 in complexes with ATF-2 in immunoprecipitates

obtained with antibodies specific for ATF-2 were quantified by western blotting. Columns 1–3, 10^{-3} , 5×10^{-4} and 10^{-4} dilutions of p300-specific antiserum; columns 4–6, 10^{-4} , 5×10^{-5} and 10^{-5} dilutions of ATF-2-specific antiserum. Results shown in all figures represent the s.e. of five experiments.

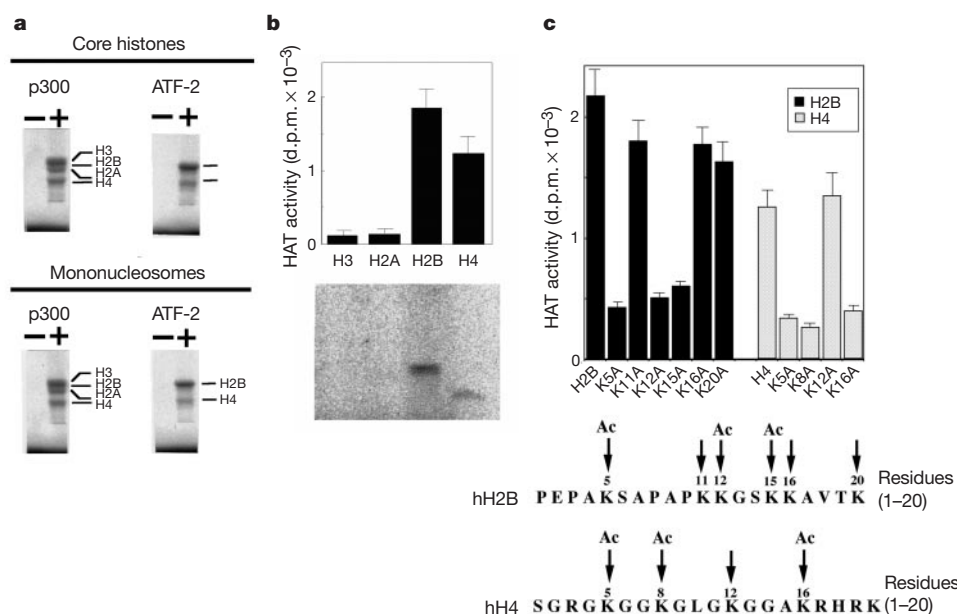


Figure 2 ATF-2 selectively acetylated human histones H2B and H4 *in vitro*. **a**, Core histones (33 μ g, upper panel) or mononucleosomes (3 μ g, lower panel) were incubated with no enzyme (–), with GST–p300 (+) or with GST–ATF-2 (+) as indicated. **b**, Purified human histone (H3, H2B, H2A or H4, 0.1 μ g of each) was used in a filter binding assay (upper panel) and in an assay with analysis by PAGE (lower panel). **c**, GST–ATF-2

preferentially acetylates the N-terminal peptide tails of human histones H2B and H4. Acetylation was assayed *in vitro* in HAT buffer^{5–7} that contained GST–ATF-2, histone H2B or H4, or a mutated peptide (with lysine changed to alanine) of H2B or H4, respectively. Acetylation was assessed by a filter binding assay.

with p300-specific or ATF-2-specific antibodies. Incubation of HeLa cell extracts with antibodies against p300 or ATF-2 immunoprecipitated the respective factors that specifically acetylated histones, but not bovine serum albumin (Fig. 1b). The immunoprecipitates obtained with antibodies against c-Jun or p21 had no HAT activity. It therefore appeared that ATF-2 might have HAT activity. We quantified the p300 coupled to ATF-2 in immunoprecipitates¹⁴ obtained with ATF-2-specific antibodies. Western blots showed that the amount of p300 associated with ATF-2 in immunoprecipitates was quite low (Fig. 1b, lanes 5,6), again indicating that ATF-2 may have intrinsic HAT activity.

Assuming that nucleosomes would be the targets for HAT activity *in vivo*, we examined the purified ATF-2 protein for its ability to acetylate free histones and histones in mononucleosomes. ATF-2 preferentially acetylated human histones H2B and H4, both as free histones and as histones bound within mononucleosomes (Fig. 2a), but did not acetylate histones H3 or H2A (Fig. 2a, b). To identify the ATF-2-specific acetylation sites within core histones, we synthesized mutant peptides in which lysines were replaced with alanine, corresponding to the amino-terminal regions of human histones. We then monitored the acetylation of each peptide by ATF-2 *in vitro*. As expected, we detected no acetylation by ATF-2 of peptides with the mutations K5A, K12A and K15A in H2B, and K5A, K8A and K16A in H4 (Fig. 2c). However, other peptides, such as those

that included K11A, K16A, and K20A mutations in H2B and the K12A mutation in H4, were acetylated by ATF-2 (Fig. 2c). These results indicate that, *in vitro*, ATF-2 can acetylate lysines at positions 5, 12 and 15 in H2B, and positions 5, 8 and 16 in H4. The substrate specificity of the acetylation of human H2B and H4 by ATF-2, as determined *in vitro*, was consistent with data obtained *in vivo*¹⁶, except that the lysine at position 12 in H4 is acetylated *in vivo* only.

To identify the region of ATF-2 responsible for its apparent HAT activity, we expressed portions of ATF-2 as GST-fusion proteins in bacteria. A filter binding assay, using equimolar amounts (40 pmol) of these purified GST-fusion proteins, determined that both ATF-2ΔN and ATF-2Δ21 had strong HAT activity. The HAT activity was specific to ATF-2, because GST itself and the deletion mutants ATF-2Δ9 and ATF-2Δ (residues 112–350) had no significant activity (Fig. 3a). Thus, the middle portion of ATF-2 (residues 112–350) appears to have intrinsic HAT activity. Comparison of the HAT domains of ATF-2 and PCAF revealed some similarity in amino-acid sequence, particularly in motif A, which binds to acetyl coenzyme A (CoA). This motif is also found in an N-acetyltransferase (N-AT) and in members of the PCAF-related family of HATs, such as GCN5 (20–30% homology; Fig. 3b)^{17,18}.

We introduced three point mutations (K296A, G297A and G299A) into the putative acetyl CoA-binding site of ATF-2, which corresponds to the site conserved in the PCAF-related family of

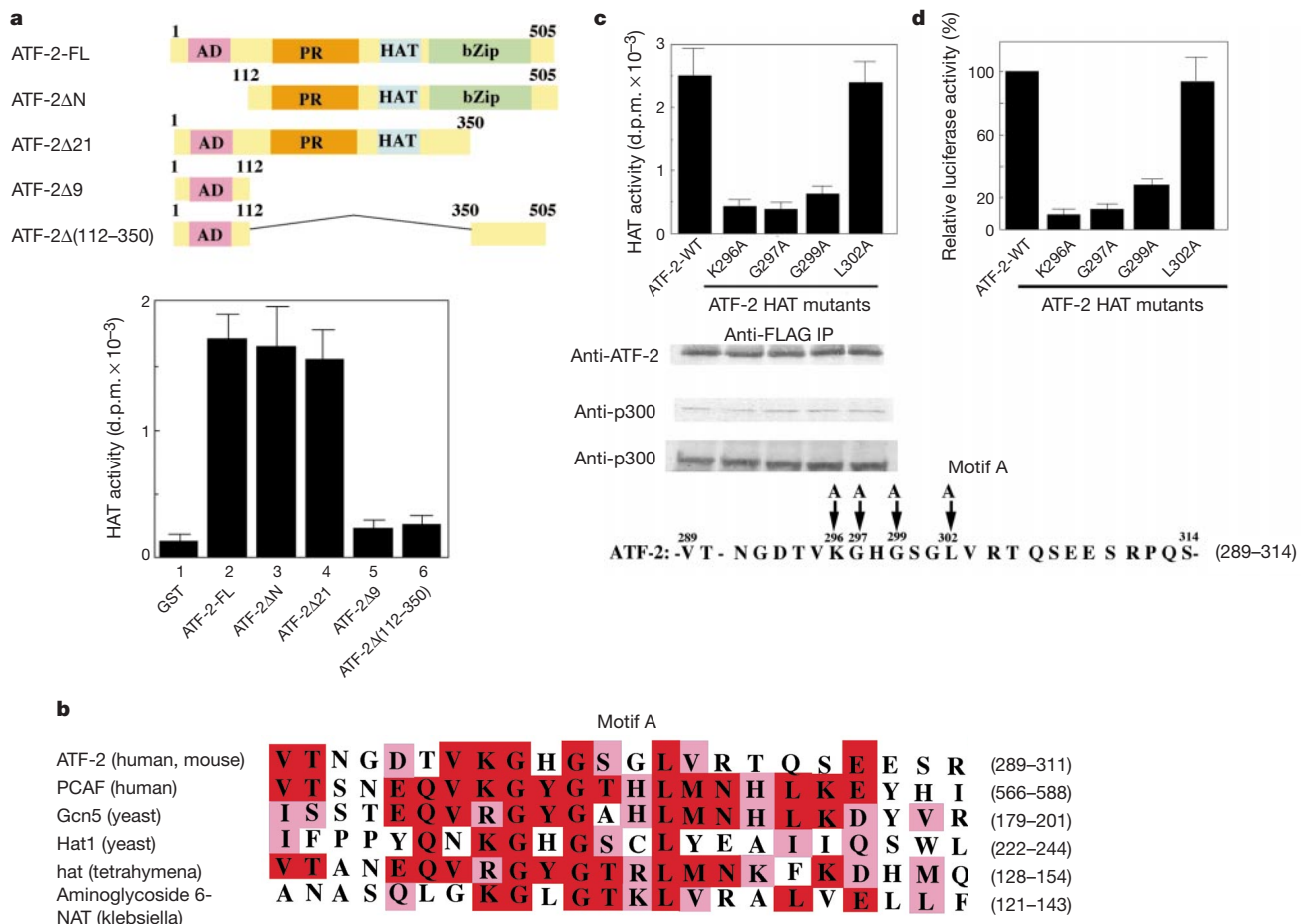


Figure 3 The proline-rich domain of ATF-2 has intrinsic HAT activity. **a**, Filter binding assay of the HAT activity of GST–ATF-2 and mutant fusion proteins (40 pmol) *in vitro*. FL, full-length; AD, activation domain; PR, proline-rich domain; HAT, histone acetyltransferase domain; bZip, basic leucine zipper domain. **b**, Sequence alignment of ATF-2 and members of the PCAF family¹⁸. Red and pink boxes indicate the same and related amino-acids residues, respectively. **c**, pCI-FLAG–ATF-2 and its mutant derivatives were used to transfect HeLa cells. Immunoprecipitation and western blotting with antibodies specific for

the flag-tag, ATF-2 or p300 were carried out. Standard immunoblotting was done with p300-specific antibodies. The flag-tag specific immunoprecipitates were tested for HAT activity *in vitro*. **d**, Stably transfected HeLa cells with a CRE-luciferase construct, were co-transfected with pCI-FLAG–ATF-2 or the same plasmid with a mutation in the gene for the HAT domain of ATF-2. Luciferase activities of pCI-FLAG–ATF-2 were arbitrarily assigned the value of 100% (refs 15, 28).

HATs¹⁷. Western blots of flag-tagged proteins immunoprecipitated by antibodies against ATF-2 (ref. 14) or p300 (ref. 19) showed that each mutant protein was expressed at the same level as wild-type ATF-2. We also found that similar levels of p300 were associated with ATF-2 in each immunoprecipitate (Fig. 3c). The three point mutations in the HAT domain each reduced the HAT activity of immunoprecipitates of ATF-2 about fivefold; however, a control derivative of ATF-2, with a mutation in a non-conserved amino acid in the HAT domain, had no reduction in HAT activity (Fig. 3c). Thus, the residues in ATF-2 that are required for acetylation of histones are localized to positions 296, 297 and 299.

We then examined the correlation between the ability of the various mutant proteins to activate CRE-dependent transcription and to acetylate histones. We detected reduced CRE-luciferase activity for the K296A, G297A and G299A mutants, but not for the control mutant (Fig. 3d). These results indicate that residues in motif A, in particular the conserved amino acids at positions 296, 297 and 299, are essential both for HAT activity and for transactivation.

Ultraviolet irradiation¹⁹ induces the phosphorylation of threonine residues T69 and T71 of ATF-2 by members of a family of stress-activated protein kinases (SAPKs)^{21,22}. Western blotting analysis of flag-specific immunoprecipitates from irradiated and non-

irradiated HeLa cells revealed a more slowly migrating band of ATF-2 on the gel after exposure to ultraviolet irradiation (Fig. 4a). This band represents a hyperphosphorylated form of ATF-2 (ref. 14). The intracellular level of ATF-2 protein was the same before and after exposure of cells to ultraviolet irradiation. Treatment of a lysate of transfected HeLa cells with calf intestinal alkaline phosphatase (CIAP) converted the slowly migrating band to a more rapidly migrating band of ATF-2. There was no significant change in the migration of ATF-2 when a group of phosphatase inhibitors was included in the reaction mixture. Thus, ultraviolet irradiation of HeLa cells apparently induces the phosphorylation of the ATF-2 protein. Results of an 'in-gel HAT assay'⁴ revealed that hyperphosphorylated ATF-2 has stronger HAT activity than unphosphorylated ATF-2 (Fig. 4a).

To confirm our observations *in vitro*, we prepared recombinant wild-type ATF-2 and mutant ATF-2 in which threonine residues T69 and T71, which are putative sites of phosphorylation by JNK/SPAK^{21,22}, were replaced by alanine. Similar mutation of the serine residue at position 90 served as a control. Recombinant ATF-2 proteins were phosphorylated by JNK *in vitro*, and their respective HAT activities were determined by an in-gel HAT assay (Fig. 4b). As expected, phosphorylated ATF-2 was more active than unphos-

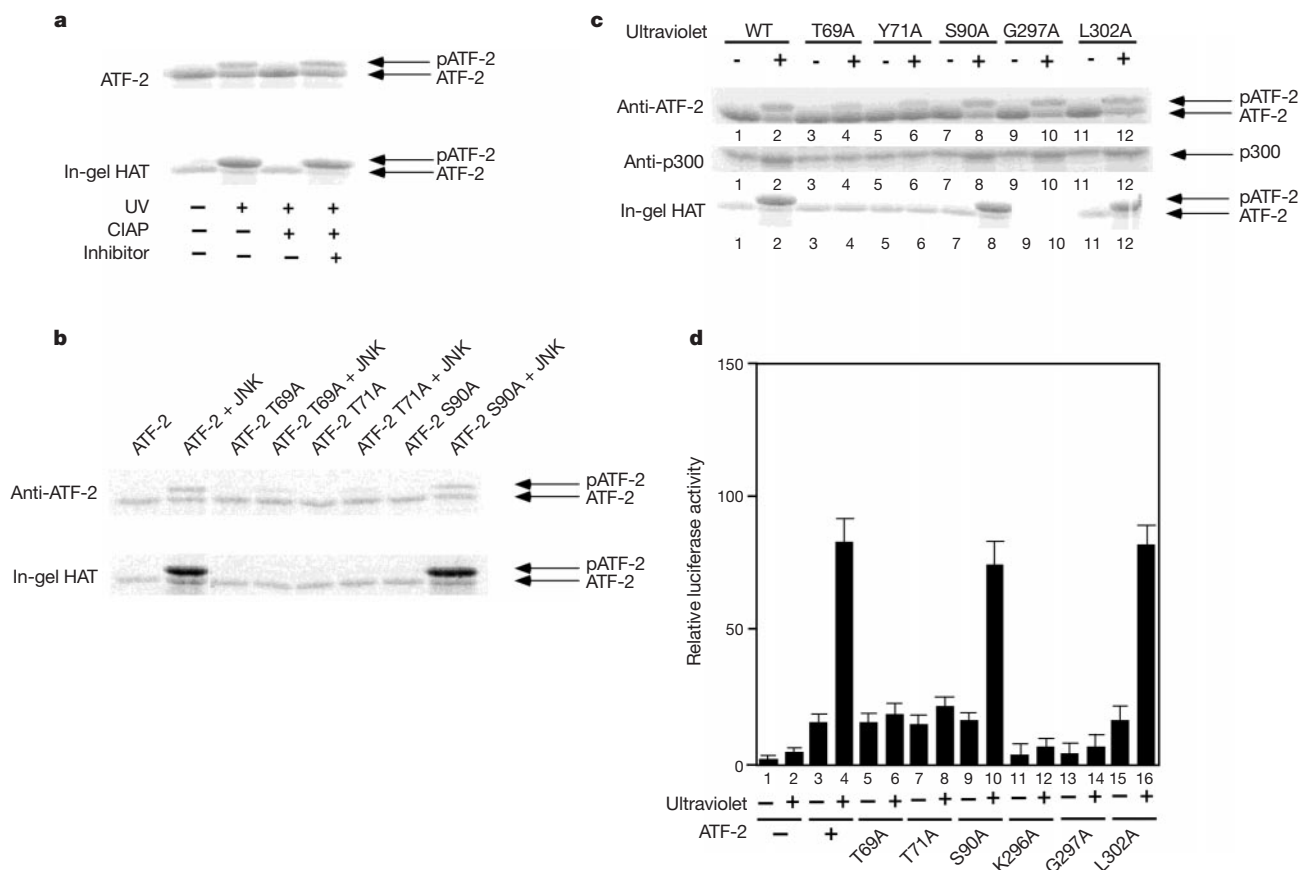


Figure 4 Phosphorylation of ATF-2 enhances its intrinsic HAT activity and CRE-dependent transcription. **a**, Lysates of HeLa cells (UV, -) or of HeLa cells after ultraviolet irradiation (UV, +) were incubated with or without calf intestinal alkaline phosphatase (CIAP) in the presence or absence of phosphatase inhibitors. The proteins were resolved by SDS-PAGE (6.5% polyacrylamide) and allowed to react with antibodies against ATF-2 after blotting (lanes 1–4). The HAT activities of ATF-2 were visualized by 'in-gel HAT assays' (lanes 5–8). The positions of ATF-2 and phosphorylated ATF-2 (pATF-2) are indicated by arrows. **b**, Recombinant ATF-2 and its derivatives (40 pmol) were incubated with JNK for 45 min at 30 °C. Phosphorylated ATF-2 was resolved by SDS-PAGE (6.5% polyacrylamide) and, after blotting, allowed to react with antibodies specific for ATF-2 (upper panel). The HAT activities of ATF-2 and its mutant derivatives were examined by

the in-gel HAT assay (lower panel)⁴. **c**, Cells that had been stably transfected with plasmids that encoded flag-tagged ATF-2 and mutant ATF-2 were incubated for 6 h with or without irradiation. Cells lysates were prepared, and western blotting analysis with antibodies specific for ATF-2 or p300 (RW128) of flag-specific immunoprecipitates was carried out after SDS-PAGE (6.5% and 5% polyacrylamide, respectively). HAT activity of ATF-2 was visualized by the in-gel HAT assay. **d**, HeLa cells that had been transfected with a plasmid that encoded CRE-luciferase were co-transfected with plasmids that encoded ATF-2, or its mutated derivatives. Transformed cells were incubated for 6 h with or without ultraviolet irradiation. Luciferase activities of pCI-FLAG were arbitrarily assigned the value of 1.0 (refs 15, 28).

phorylated ATF-2. In addition, the ATF-2 mutants T69A and T71A showed no phosphorylation-related changes in HAT activity. The control ATF-2 mutant S90A did not have reduced HAT activity. These results suggest that phosphorylation of ATF-2 at T69 or T71 may control the intrinsic HAT activity.

To determine whether phosphorylation of ATF-2 at positions 69 and 71 is required for activation of the HAT activity of ATF-2 *in vivo*, we examined levels of ATF-2 by immunoprecipitation and western blotting analysis. Expression levels of wild-type and mutant ATF-2 were unchanged by exposure of cells to ultraviolet irradiation (Fig. 4c, upper panel). The slowly migrating band of ATF-2 was detected after exposure of cells to ultraviolet irradiation in the case of wild-type ATF-2, S90A, a protein with a mutation in the HAT domain (G297A) and the control mutant protein (L302A). We then examined the amounts of p300 associated with ATF-2. An intense band corresponding to p300 was detected in the case of transfectants that expressed wild-type ATF-2, S90A, G297A or L302A after exposure of cells to ultraviolet light (middle panel). In contrast, p300 bands were weak in the case of HeLa cells that expressed T69A or T71A. Other HAT domain mutants, such as K296A and G299A, gave similar results to the G297A mutant (data not shown). Ultraviolet-induced ATF-2 phosphorylation resulted in increased HAT activity, compared with unphosphorylated ATF-2 (lower panel). T69A and T71A had weak HAT activity, whereas G297A had no HAT activity either before or after ultraviolet irradiation. Similar results were obtained with two other HAT domain mutants, K296A and G299A (data not shown). The HAT activity of L302A was similar to that of wild-type ATF-2. These data strongly suggest that the HAT activity of ATF-2 is enhanced by phosphorylation at T69 or T71. Moreover, the conserved amino acids in the HAT domain at positions 296, 297 and 299 appear to have a critical role in the HAT activity of ATF-2 after exposure of cells to ultraviolet irradiation.

To investigate the effects of phosphorylation of ATF-2 on transcription of a CRE-luciferase reporter gene, we generated stably transformed HeLa cells that harboured a reporter gene encoding CRE-luciferase. Co-transfection with wild-type ATF-2 and exposure of cells to ultraviolet light increased the transcription of the CRE-luciferase reporter gene, as compared with in the absence of such exposure (Fig. 4d). Mutation of ATF-2 at T69 or T71, but not S90A, destroyed the promoting effect of ATF-2. Thus, phosphorylation of ATF-2 at T69 or T71 is required for the HAT activity of ATF-2 and for stimulation by ATF-2 of CRE-mediated transcription after exposure of ultraviolet irradiation. Notably, no stimulation of transcription of a CRE-luciferase reporter gene was detected in the case of ATF-2 proteins with mutations in the HAT domain (Fig. 4d). Similar results were obtained with G299A (data not shown). The control mutant protein L302A stimulated transcription of the CRE reporter to a similar degree as the wild type. These results suggest that the HAT domain of ATF-2 is critical for enhanced transcription of the CRE reporter gene in transfected HeLa cells after ultraviolet irradiation.

We have shown that ATF-2, a DNA-binding protein, has intrinsic HAT activity with unique specificity for histones H2B and H4 by experiments involving *in vitro* phosphorylation by JNK and an in-gel HAT assay using phosphorylated ATF-2 proteins (Fig. 4b). Our findings suggest that sequence-specific transcription factors target specific promoters and then recruit various HATs to these specific promoters to enhance transcription. Moreover, we have shown that a sequence-specific factor, ATF-2, has intrinsic HAT activity; ATF-2 might also recruit other HATs that are required for transcriptional activation. The phosphorylation of ATF-2 appears to have a key role in stimulating the HAT activity of ATF-2 and in promoting CRE-dependent transcription. The HAT domain of ATF-2 has substantial homology to motif A of HATs in the PCAF family, and the conserved amino acids at 296, 297 and 299 in motif A are critical for the intrinsic HAT activity (Fig. 3c, d).

Why does the phosphorylation, by JNK/SPAK, of the N-terminal

region of ATF-2 outside the HAT domain affect the intrinsic HAT activity? The ability of the N-terminal region of ATF-2 to function as a transcription-activation domain is correlated with its involvement in mediating stimulation by E1A. The phosphorylation of T69 and T71 is essential for this process^{15,22}. These residues are phosphorylated *in vitro* by members of the JNK/SPAK family²², and the phosphorylation of the N-terminal domain can be stimulated by exposure of cells to ultraviolet irradiation (Fig. 4c). The phosphorylation induced by JNK might induce a conformational change that exposes the binding site of acetyl-CoA or the catalytic domain of ATF-2, thereby activating the HAT. The specificity of the histone acetylation catalysed by ATF-2 is also intriguing. The H2B–H4 interface is a likely site for initial disruption of histone–histone interactions upon unfolding of the nucleosome *in vitro* and *in vivo*²³. It is possible that ATF-2, a specific DNA-binding factor, might have a key role in the stability of the H2B–H4 interaction by acetylation and/or by direct binding to a specific site on DNA. However, we cannot rule out the possibility that non-histone proteins, such as p53 (ref. 24), EKLF (ref. 25), GATA-1 (ref. 26), TFIIE- β (ref. 10) and TFIIF (ref. 10), might be targets of the HAT activity of ATF-2. The functional roles of two distinct HAT activities, namely, the intrinsic activity of ATF-2 itself and the HAT activity associated with the p300 coactivator, merit further investigation. $\square$

Methods

Materials and plasmids

Chemically synthesized peptides and variants with alanine residues replacing lysines, corresponding to the N-terminal regions of human histones H2B and H4, were obtained from Biologica. Complementary DNAs for GST–ATF-2 and His–ATF-2 derivatives, cDNAs for expression of flag-tagged p300, ATF-2 deletion mutants of the HAT domain and pCI-FLAG vectors of ATF-2 derivatives were generated as described^{5,7,14,15}. Other cDNAs for c-Jun, ATF-2, JNK and p38 were gifts from M. Karin. Histidine-tagged fusion proteins and GST-fusion proteins were affinity purified with glutathione-conjugated agarose or Ni-NTA agarose beads^{14,22}.

Transfection, ultraviolet irradiation and assay of luciferase activity

HeLa S3 cells were transfected with LipofectAMINE 2000 (GIBCO-BRL)²⁷, according to the manufacturer's protocol, with 10 μ g of CRE-luciferase plasmid²⁸ and pSV- β -Gal (ref. 14). The G418-resistant stable clones (5×10^6) were selected and transfected again with 10 μ g of expression plasmids and 8 μ g of pRSVhygro, as well as 2 μ g of pSV- β -Gal (ref. 14), and cultured in the presence of hygromycin for two weeks. A pool of G418- and hygromycin-resistant cells was irradiated with light from a monochromatic ultraviolet lamp (15 W Hg lamp) at 254 nm (half-maximal bandwidth, 2.3 nm; 0.1 mW cm⁻², 60–100 J m⁻², 6 h) as described²⁹. Then cells were harvested, lysed and assayed for luciferase activity as described²⁷.

Immunoprecipitation and immunoblotting

Preparation of cell lysates, immunoprecipitation with ATF-2-specific polyclonal or monoclonal antibodies, and treatment with calf intestinal alkaline phosphatase (CIAP) were done as described¹⁴. Immunoprecipitated proteins were separated by SDS–PAGE (6.5% polyacrylamide) and visualized by immunoblotting¹⁴ with ATF-2-specific antibodies (UB; Santa Cruz Biotechnology)², p300-specific antibodies (RW102 or RW128)¹⁹, or M2 flag-specific antibodies (Kodak-IBI).

Assay of HAT activity

Filter binding assays of HAT activity were done as described^{5,7}. In some cases, 2 pmol of enzyme protein, 1 μ g of substrate, 90 pmol of [¹⁴C]acetyl CoA (5.8 Ci mmol⁻¹; Amersham Pharmacia) and 9 pmol of mononucleosomes were used. Assays of immunoprecipitated HAT activity were done as described^{5–8}. We used polyclonal antibodies against c-Jun (Santa Cruz), polyclonal antibodies against cdk inhibitor p21^{Cip1} (Santa Cruz) and monoclonal antibodies against p38 (Santa Cruz).

Phosphorylation of ATF-2 *in vitro*

Recombinant ATF-2 (40 pmol) was incubated with 20 ng of JNK (UB) or p38 (UB) plus 100 μ M ATP (including [γ -³²P]ATP to give a final specific activity of 100 mCi mmol⁻¹ to measure the specific kinase activity, where indicated) for 45 min at 30 °C in 30 μ l of 25 mM Tris–HCl pH 7.5, 0.1 mM Na₂VO₄, 0.1 mM EGTA, 10 mM magnesium acetate, 0.04 mM DTT, 0.1 mM ZnSO₄, supplemented with protease inhibitors as described²². Reaction mixtures were fractionated by SDS–PAGE (6.5% polyacrylamide) and western blotting with antibodies specific for ATF-2 was carried out as indicated.

In-gel HAT assay⁴

Preparations containing ATF-2 were separated by SDS–PAGE on a 6.5% polyacrylamide

gel containing 1 mg ml⁻¹ core histones. The gel was washed with 50 mM HEPES pH 7.9, 25% isopropanol for 30 min at room temperature. The proteins were denatured by incubation with 6 M guanidine-HCl in 50 mM HEPES pH 7.9, 1 mM DTT, 5 mM PMSF for 30 min at room temperature. The proteins on the gel were renatured by incubation with chilled 50 mM HEPES pH 7.9, 1 mM DTT, 5 mM PMSF and 0.05% Tween 20 at 4 °C for 12 h. After renaturation of ATF-2 in the gel, the gel was incubated with [¹⁴C]acetyl CoA in reaction buffer (50 mM HEPES pH 7.9, 1 mM sodium butyrate, 1 mM DTT, 5 mM PMSF and 2.5 mM MgCl₂) at 30 °C for 1 h. Acetylated histones were visualized by autoradiography.

Received 8 December 1999; accepted 22 February 2000.

- Hai, T., Liu, F., Coukos, W. J. & Green, M. R. Transcription factor ATF cDNA clones: an extensive family of leucine zipper proteins able to selectively form DNA-binding heterodimers. *Genes Dev.* **3**, 2083–2090 (1989).
- Maekawa, T. *et al.* Leucine zipper structure of the protein CRE-BP 1 binding to the cyclic AMP-response element in brain. *EMBO J.* **8**, 2023–2028 (1989).
- Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386**, 569–573 (1997).
- Brownell, J. E. *et al.* Tetrahymena histone acetyltransferase A: a homolog of yeast Gcn5p linking histone acetylation to gene activation. *Cell* **84**, 843–851 (1996).
- Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87**, 953–959 (1996).
- Bannister, A. J. & Kouzarides, T. The CBP co-activator is a histone acetyltransferase. *Nature* **384**, 641–643 (1996).
- Yang X. J., Ogryzko, V. V., Nishikawa, J., Howard, B. H. & Nakatani, Y. A p300/CBP-associated factor that competes with the adenoviral oncoprotein E1A. *Nature* **382**, 319–324 (1996).
- Chen, H. *et al.* Nuclear receptor coactivator ACTR is a novel histone acetyltransferase and forms a multimeric activation complex with P/CAF and CBP/p300. *Cell* **90**, 569–580 (1997).
- Spencer, T. E. *et al.* Steroid receptor coactivator-1 is a histone acetyltransferase. *Nature* **389**, 194–198 (1997).
- Imhof, A. *et al.* Acetylation of general transcription factors by histone acetyltransferases. *Curr. Biol.* **7**, 689–692 (1997).
- Torchia, J. *et al.* The transcriptional co-activator p/CIP binds CBP and mediates nuclear-receptor function. *Nature* **387**, 677–684 (1997).
- Struhl, K. Histone acetylation and transcriptional regulatory mechanisms. *Genes Dev.* **12**, 599–606 (1998).
- Brown, C. E., Lechner, T., Howe, L. & Workman, J. L. The many HATs of transcription coactivators. *Trends Biochem. Sci.* **25**, 15–19 (2000).
- Kawasaki, H. *et al.* p300 and ATF-2 are components of the DRF complex, which regulates retinoic acid- and E1A-mediated transcription of the *c-jun* gene in F9 cells. *Genes Dev.* **12**, 233–245 (1998).
- Liu, F. & Green, M. A specific member of the ATF transcription factor family can mediate transcription activation by the adenovirus E1A protein. *Cell* **61**, 1217–1224 (1990).
- Schiltz, R. L. *et al.* Overlapping but distinct patterns of histone acetylation by the human coactivators p300 and P/CAF within nucleosome substrates. *J. Biol. Chem.* **274**, 1189–1192 (1999).
- Martinez-Balbas, M. A. *et al.* The acetyltransferase activity of CBP stimulates transcription. *EMBO J.* **17**, 2886–2893 (1998).
- Dutnall, R. N., Tafrov, S. T., Sternglanz, R. & Ramakrishnam, V. Structure of the histone acetyltransferase Hat1: a paradigm for the GCN5-related N-acetyltransferase superfamily. *Cell* **94**, 427–438 (1998).
- Eckner, R., Yao, T. -S., Oldread, E. & Livingston, D. M. Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kDa protein (p300) reveals a protein with properties of a transcriptional adaptor. *Genes Dev.* **8**, 869–884 (1994).
- Herrlich, P., Blattner, C., Knebel, A., Bender, K. & Rahmsdorf, H. J. Nuclear and non-nuclear targets of genotoxic agents in the induction of gene expression. Shared principles in yeast, rodents, man and plants. *Biol. Chem.* **378**, 1217–1229 (1997).
- van Dam, H. *et al.* ATF-2 is preferentially activated by stress-activated protein kinases to mediate *c-jun* induction in response to genotoxic agents. *EMBO J.* **14**, 1798–1811 (1995).
- Livingston, C., Patel, G. & Jones, N. C. ATF-2 contains a phosphorylation-dependent transcriptional activation domain. *EMBO J.* **14**, 1785–1797 (1995).
- Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. X-ray structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–259 (1997).
- Gu, W. & Roeder, R. G. Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain. *Cell* **90**, 595–606 (1997).
- Zhang, W. & Bieker, J. J. Acetylation and modulation of erythroid Krüppel-like factor (EKLF) activity by interaction with histone acetyltransferases. *Proc. Natl Acad. Sci. USA* **95**, 9855–9860 (1998).
- Boyes, J., Byfield, P., Nakatani, Y. & Ogryzko, V. V. Regulation of activity of the transcription factor GATA-1 by acetylation. *Nature* **396**, 594–598 (1998).
- Kawasaki, H. *et al.* Distinct roles of the co-activators p300 and CBP in retinoic acid-induced F9-cell differentiation. *Nature* **393**, 284–289 (1998).
- Montminy, M. R. & Beizikjian, J. M. Binding of a nuclear protein to the cyclic-AMP response element of the somatostatin gene. *Nature* **328**, 175–178 (1987).
- Bender, K., Gottlicher, M., Whiteside, S., Rahmsdorf, H. J. & Herrlich, P. Sequential DNA damage-independent and -dependent activation of NF-κB by ultraviolet. *EMBO J.* **17**, 5710–5718 (1998).

Acknowledgements

We thank T. Nakajima, S. Ishii, N. Jones, M. Green and M. Karin for providing reagents; C. Geltinger, G. Gachelin, T.-P. Yao, D. M. Livingston, A. P. Wolffe and R. Eckner for discussions; and N. Day, A. Körner and B. C. An der Lan for critical reading of the manuscript. This work was supported by the NIH (R.C., Y.N., K.I.), by the Special Cooperation Funds of the Science and Technology Agency, and by Grants-in-Aid from the Ministry of Education, Science, Sports and Culture of Japan (K.K.Y.).

Correspondence and requests for materials should be addressed to K.K.Y. (e-mail: kazunari@rtc.riken.go.jp).

B and C floral organ identity functions require *SEPALLATA* MADS-box genes

Soraya Pelaz*, Gary S. Ditta*, Elvira Baumann†, Ellen Wisman† & Martin F. Yanofsky*

* Section of Cell and Developmental Biology, La Jolla, California 92093-0116, USA

† Max-Planck-Institut für Züchtungsforschung, Carl-von-Linné-Weg 10, 50829 Köln, Germany

Abnormal flowers have been recognized for thousands of years, but only in the past decade have the mysteries of flower development begun to unfold. Among these mysteries is the differentiation of four distinct organ types (sepals, petals, stamens and carpels), each of which may be a modified leaf¹. A landmark accomplishment in plant developmental biology is the ABC model of flower organ identity^{2,3}. This simple model provides a conceptual framework for explaining how the individual and combined activities of the ABC genes produce the four organ types of the typical eudicot flower. Here we show that the activities of the B and C organ-identity genes require the activities of three closely related and functionally redundant MADS-box genes, *SEPALLATA1/2/3* (*SEP1/2/3*). Triple mutant *Arabidopsis* plants lacking the activity of all three *SEP* genes produce flowers in which all organs develop as sepals. Thus *SEP1/2/3* are a class of organ-identity genes that is required for development of petals, stamens and carpels.

The field of flower development has truly blossomed during the past decade, with the identification of numerous mutants and the corresponding genes. One class of genes that has attracted considerable attention is the ABC organ-identity genes, mutations in which lead to homeotic conversions of organ types, indicating that the ABC genes specify the fate of flower organ primordia^{2–4}. The ABC genes act alone or in combination to specify the four flower organ types (Fig. 1a). One of the principles of the ABC model is that each activity is confined to two adjacent whorls, allowing three activities to specify the four organ types of a typical eudicot flower.

Central to the ABC model is a single family of DNA-binding transcriptional regulators, encoded by the MADS-box genes. These genes encode key components of the A, B and C activities, and the expression of these genes is largely confined to the whorls where they specify organ identity (Fig. 1b). The A-function gene *APETALA1* (*API*) is expressed in the two outer whorls by stage 3 of flower development (stages defined in ref. 5), when sepal primordia are just becoming visible^{6,7}. The B-function genes *APETALA3* (*AP3*) and *PISTILLATA* (*PI*) are expressed in whorls two and three^{8,9}, and the C-function gene *AGAMOUS* (*AG*) is expressed in the two inner whorls¹⁰. In addition to the MADS-box genes, other genes contribute to A activity. Most notable is *APETALA2* (*AP2*), which encodes a putative transcription factor that is necessary to prevent *AG* RNA from accumulating in the two outer whorls^{5,11}. However, *AG* is the only known C-function gene, and *AP3* and *PI* are the only known B-function genes. In spite of the remarkable progress in understanding these fundamental steps in flower development, only a small part of the overall picture has emerged. The importance of MADS-box genes cannot be overstated as they control flowering time, meristem identity and key components of the ABC organ-identity model.

The genes controlling flower development have mostly been defined through traditional forward mutagenesis screens. However, there are many examples of functional redundancy among closely related genes^{12–14}, indicating that a reverse-genetics approach may be needed to identify genes that would have gone unnoticed through traditional approaches. MADS-box genes control diverse

aspects of plant development and their functional roles correlate closely with their domains of RNA accumulation, so it is probable that other MADS-box genes expressed early in flower development also have important functions. Among these are *AGAMOUS-LIKE* (*AGL*)2, *AGL*4 and *AGL*9 (refs 15, 16), which are first expressed just before the onset of *B* and *C* MADS-box gene expression. *AGL*2, *AGL*4 and *AGL*9 are all expressed in the three inner whorls of organs throughout flower development, although *AGL*2 and *AGL*4 are also expressed in the outermost whorl of young flowers^{17,18}. We have renamed the *AGL*2, *AGL*4 and *AGL*9 genes *SEPALLATA*1 (*SEP*1), *SEP*2 and *SEP*3, respectively, because the triple mutant flower consists of 'lots of sepals' (Fig. 2).

To discover the roles of *SEP*1/2/3 during flower development, we used a reverse-genetics polymerase chain reaction (PCR)-based approach to identify mutant alleles. We found that single gene mutations produced only subtle phenotypes, consistent with the possibility of functional redundancy (unpublished results). However, a striking phenotype occurred in *sep1 sep2 sep3* triple mutants, in which all flower organs resembled sepals (Fig. 2). The second whorl, which is normally occupied by four white petals (Fig. 2a), has four green sepal-like organs instead (Fig. 2b, c). Close inspection of the converted second whorl organs by scanning electron microscopy revealed the differentiation of interspersed stomata and the conversion of nearly all of the cells in these organs into sepal cells, particularly on the abaxial (outer) surface (Fig. 2g). The third whorl is normally occupied by six stamens, each consisting of a long filament capped by a pollen-bearing anther. However in the triple mutant they are green sepal-like organs that have differentiated cell types characteristic of sepals, including stomata (Fig. 2h). The fourth whorl, which normally consists of two fused carpels, is a new flower that reiterates the same basic pattern (sepal, sepal, sepal)¹⁹. Thus, whereas normal flowers are determinate structures consisting of only four whorls of organs, the triple mutant flowers

are indeterminate and produce a seemingly endless reiteration of flower organs. These studies show that the *SEP*1/2/3 genes have overlapping functions required for petal, stamen and carpel development. In addition, these three genes are required to prevent the indeterminate growth of the flower meristem.

The *sep1/2/3* triple mutant phenotype is strikingly similar to *bc* (*ap3 ag, pi ag*) double mutants³ (Fig. 2d), indicating that the *B* and *C* organ-identity genes are inactive in the triple mutant. These data indicate that *SEP*1/2/3 are required to activate the *B* and *C* genes, or that the *B* and *C* gene products require at least one of the *SEP*1/2/3 proteins for activity. The observation that *SEP*1, *SEP*2 and *SEP*3 are still expressed in *b* and *c* organ-identity mutants indicates that they probably do not act downstream of the *B* and *C* genes. To determine whether the *SEP*1/2/3 genes are required for *B* and *C* gene activation, we analysed the expression of *AP*3, *PI* and *AG* in *sep1/2/3* triple mutant flowers. The initial patterns of *B* and *C* expression were not altered in the triple mutant (Fig. 3), indicating that the *SEP*1/2/3 genes are not required for the initial activation of the *B* and *C* genes.

On the basis of these results, we propose that *SEP*1/2/3 encode a new type of organ-identity activity that encompasses both *B* and *C*

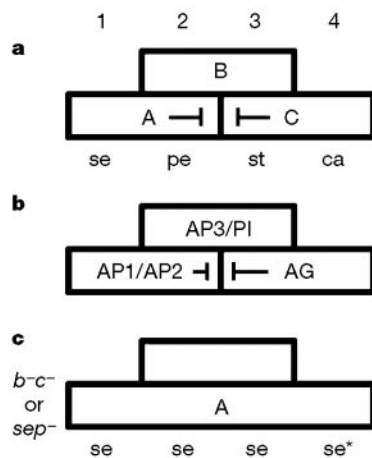


Figure 1 Models for flower organ identity in *Arabidopsis*. **a**, The ABC model of flower organ identity: A, B and C activities are each active in two adjacent whorls where their individual and combined activities specify the fate of organ primordia. A alone (whorl 1) specifies sepals, C alone (whorl 4) specifies carpels, and the combined activities of AB (whorl 2) and BC (whorl 3) specify petals and stamens, respectively. The A and C activities are mutually antagonistic, such that A prevents the activity of C in the outer two whorls, and C prevents the activity of A in the inner two whorls. The C activity is also required for floral determinacy: if the C activity is absent, an indeterminate number of floral whorls develop (asterisk in **c**). **b**, In *Arabidopsis* the A activity includes *AP*1 and *AP*2, B includes *AP*3 and *PI*, and the only C-function gene is *AG*. With the exception of the ubiquitously expressed *AP*2 gene, all are MADS-box genes that are expressed in the regions where they specify organ identity. **c**, In a typical *bc* double mutant (*ap3 ag* or *pi ag*), the only remaining activity is A which expands to include all whorls. These double mutants have the phenotype (sepal, sepal, sepal)¹⁹. The same phenotype occurs in the *sep1 sep2 sep3* triple mutant, indicating that B and C functions are inactive.

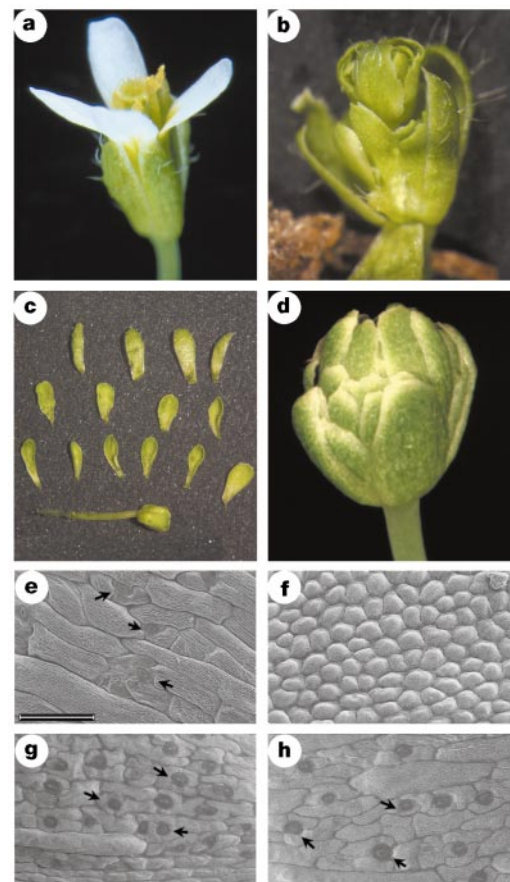


Figure 2 Phenotype of *sep1 sep2 sep3* triple mutants. **a**, Wild-type flower consisting of four sepals, four petals, six stamens and two fused carpels. **b**, *sep1 sep2 sep3* triple mutant flower in which the four petals and six stamens are replaced by sepaloid organs and carpels are replaced by a new flower that repeats this same phenotype. In addition, there is internode elongation between internal flowers, presumably because of a functional *ERECTA* gene. **c**, Dissected *sep1 sep2 sep3* triple mutant flower with first-whorl sepals (top), second and third whorl sepaloid organs (middle), and a new flower (bottom) that replaces the carpels. **d**, *pi ag* (*bc*) double mutant that reiterates the same sepal, sepal, sepal phenotype. **e**, Abaxial surface of wild-type sepal. **f**, Abaxial surface of wild-type petal. **g**, Abaxial surface of *sep1 sep2 sep3* second-whorl sepaloid organ. **h**, Abaxial surface of *sep1 sep2 sep3* third-whorl sepaloid organ. Arrows indicate several stomata. Scale bar in e-h: 50 μm.

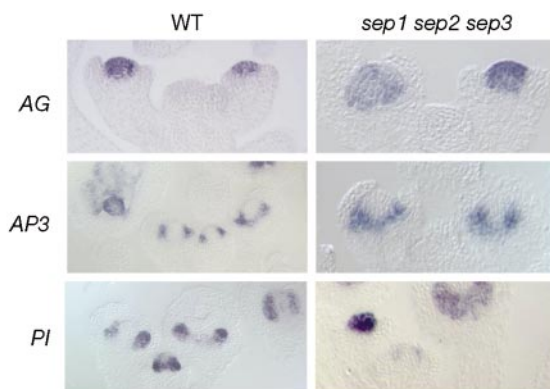


Figure 3 *AP3*, *PI* and *AG* are activated normally in *sep1 sep2 sep3* mutants. The accumulation of *AG* RNA in whorls three and four appears unchanged in *sep1 sep2 sep3* triple mutants. Similarly, the accumulation of *AP3* and *PI* RNAs in whorls two and three appears unchanged in *sep1 sep2 sep3* triple mutants. WT, wild type.

functions. In contrast to the *ABC* genes, which specify organ identity in two adjacent whorls, the activities of *SEP1/2/3* span the three inner whorls. The fact that the *SEP* genes were not included in the original *ABC* model may be an historical artefact because their redundancy hid their important organ-identity functions. The power of reverse genetics to discover proteins with highly redundant functions may lead to further refinements of the *ABC* model, in which case it might be useful to think of *SEP* as a combined B/C-class, or even a D-class gene.

One unanswered question is whether the *SEP* proteins interact directly with the products of the *ABC* genes. B function is mediated by the *AP3/PI* heterodimer^{9,19}, and yeast two-hybrid studies show that *AP1* interacts with *SEP3* and *AG* interacts with *SEP1*, *SEP2* and *SEP3* (ref. 20 and unpublished results). Furthermore, the *AP3/PI* heterodimer can interact with *SEP3* and *AP1*, as well as *SEP3* and *AG* in larger complexes (T. Homma and K. Goto, personal communication). *Antirrhinum* orthologues of these MADS-box gene products also form ternary complexes^{21,22}. The genetic and protein–protein interaction data strongly support the idea that the *SEP* proteins interact directly with the products of the *ABC* genes, and that larger complexes of MADS-box gene products may direct organ identity.

The *ABC* model of flower organ identity appears to be applicable to diverse plant species^{2,3,23} and it will be interesting to determine whether the organ-identity function of the *SEP* genes is similarly conserved. Candidate orthologues of *SEP* genes with similar expression domains exist in distantly related eudicots^{21,24,25} and in monocots²⁶, indicating that the conclusions from studies of the *SEP* genes in *Arabidopsis* will be applicable to diverse species. Indeed, antisense and cosuppressed lines of candidate *SEP3* orthologues in petunia (*FBP2*) and tomato (*TM5*) have phenotypes resembling our triple mutant²⁵. Gymnosperms also have apparent *SEP1/2* orthologues²⁷ that are expressed early in reproductive development, indicating that the importance of these genes predates the emergence of flowering plants.

The theory that flower organs are simply modified leaves is supported by the observation that triple mutants lacking all three of the *ABC* gene activities produce flowers in which all organs are leaf-like^{1,3}. However, attempts to convert vegetative leaves into each of the distinct floral organs by ectopic expression of the *ABC* genes have been largely unsuccessful^{28,29}, indicating that at least one other factor is needed for this conversion. The *SEP* genes, which are required for petal, stamen and carpel development, may be the missing factors. It will be interesting to determine whether the combined action of *SEP* and the *ABC* genes will be sufficient to convert leaves into flower organs. □

Methods

Identification of mutant alleles

To identify insertion alleles of *SEP2* and *SEP3*, we screened a population of *Arabidopsis* Col plants that carries about 50,000 independent insertions of the autonomous maize transposable element *En-1* (ref. 30). Primers used in PCR reactions were: *SEP2*-X12: 5'-AGCAAGTTCGCTGCATCAAGGTGA-3'; *SEP2*-X2: 5'-TCCAGCCAGGGATG-TAGCCGTTTCCTT-3'; *SEP3*-TT5: 5'-TCCTATGAGGGTCTTTGGTACACAATAATT-3'; and *SEP3*-X2: 5'-GCACACAAGACAGAAAACGTGAG-3'. Each oligonucleotide was used with both *En-1* specific primers En205 and En8130. We used genomic DNA from single plants to amplify the region of *En-1* insertion by PCR, and identified the insertion site by subsequent sequence analyses. We identified two different alleles of *SEP3* and one of *SEP2*. The insertion in *sep2-1* is in the seventh intron, 1,889 bases after the start codon; the 5' end of the *En-1* is towards the 5' end of *SEP2*. Based on similar mutations in other MADS-box genes, *sep2-1* probably represents an intermediate allele. The *En-1* transposon has inserted into *sep3-1*, 1,050 bases after the start codon, in the second exon, with the 5' end of the transposon towards the start. The *En-1* element has inserted into *sep3-2*, 48 bases after the start codon, in the first exon, with the 3' end towards the start of the gene. To identify an insertion allele for *SEP1*, we screened the DuPont collection of 13,800 T-DNA insertional lines with about 19,800 independent inserts. We used the TLFT oligonucleotide of the T-DNA insert (5'-GATGCACATCGAAATCAGCCAATTTAGAC-3') and the *SEP1*-5'L oligonucleotide specific for *SEP1* (5'-CACAACTCCACACACTTCCAAACAC-3'). Sequence analyses of genomic DNA identified the insertion site at 1,364 bases from the start codon in the third intron. We identified homozygous plants by genomic Southern blotting.

Expression and microscopy analyses

For scanning electron microscopy, we fixed flowers overnight in 50% ethanol, 5% acetic acid, 3.7% formaldehyde. For RNA *in situ* hybridization experiments we used digoxigenin-11-rUTP labelled probes and a nucleic acid labelling kit (Boehringer) following the manufacturer's instructions¹³. We synthesized the *AG* antisense probe by digesting pCIT565 with *HindIII* and then using T7 polymerase³. We digested PD793 with *BglII* and then used T7 polymerase to produce the *AP3* antisense probe⁸. We digested pcPINX2 (a gift from K. Goto) with *BamHI* and used T7 polymerase to obtain the *PI* antisense probe⁹.

Received 21 January; accepted 28 February 2000.

- Meyerowitz, E. M., Smyth, D. R. & Bowman, J. L. Abnormal flowers and pattern formation in floral development. *Development* **106**, 209–217 (1989).
- Coen, E. S. & Meyerowitz, E. M. The war of the whorls: genetic interactions controlling flower development. *Nature* **353**, 31–37 (1991).
- Bowman, J. L., Smyth, D. R. & Meyerowitz, E. M. Genetic interactions among floral homeotic genes of *Arabidopsis*. *Development* **112**, 1–20 (1991).
- Schwarz-Sommer, Z., Huijser, P., Nacken, W., Saedler, H. & Sommer, H. Genetic control of flower development: homeotic genes in *Antirrhinum majus*. *Science* **250**, 931–936 (1990).
- Drews, G. N., Bowman, J. L. & Meyerowitz, E. M. Negative regulation of the *Arabidopsis* homeotic gene *AGAMOUS* by the *APETALA2* product. *Cell* **65**, 991–1002 (1991).
- Irish, V. F. & Sussex, I. M. Function of the *APETALA1* gene during *Arabidopsis* floral development. *Plant Cell* **2**, 741–751 (1990).
- Mandel, M. A., Gustafson-Brown, C., Savidge, B. & Yanofsky, M. F. Molecular characterization of the *Arabidopsis* floral homeotic gene *APETALA1*. *Nature* **360**, 273–277 (1992).
- Jack, T., Brockman, L. L. & Meyerowitz, E. M. The homeotic gene *APETALA3* of *Arabidopsis thaliana* encodes a MADS-box and is expressed in petals and stamens. *Cell* **68**, 683–697 (1992).
- Goto, K. & Meyerowitz, E. M. Function and regulation of the *Arabidopsis* floral homeotic gene *PISTILLATA*. *Genes Dev.* **8**, 1548–1560 (1994).
- Yanofsky, M. F. *et al.* The protein encoded by the *Arabidopsis* homeotic gene *agamous* resembles transcription factors. *Nature* **346**, 35–39 (1990).
- Jofuku, K. D., den Boer, B. G. W., Van Montagu, M. & Okamura, J. K. Control of *Arabidopsis* flower and seed development by the homeotic gene *APETALA2*. *Plant Cell* **6**, 1211–1225 (1994).
- Kempin, S. A., Savidge, B. & Yanofsky, M. F. Molecular basis of the *cauliflower* phenotype in *Arabidopsis*. *Science* **267**, 522–525 (1995).
- Ferrández, C., Gu, Q., Martienssen, R. & Yanofsky, M. F. Redundant regulation of meristem identity and plant architecture by *FRUITFULL*, *APETALA1* and *CAULIFLOWER*. *Development* **127**, 725–734 (2000).
- Liljgren, S. J. *et al.* *SHATTERPROOF* MADS-box genes control seed dispersal in *Arabidopsis*. *Nature* **404**, 766–770 (2000).
- Ma, H., Yanofsky, M. F. & Meyerowitz, E. M. *AGL1-AGL6*, an *Arabidopsis* gene family with similarity to floral homeotic and transcription factor genes. *Genes Dev.* **5**, 484–495 (1991).
- Mandel, M. A. & Yanofsky, M. F. The *Arabidopsis* *AGL9* MADS-box gene is expressed in young flower primordia. *Sex. Plant Reprod.* **11**, 22–28 (1998).
- Flanagan, C. A. & Ma, H. Spatially and temporally regulated expression of the MADS-box gene *AGL2* in wild-type and mutant *Arabidopsis* flowers. *Plant Mol. Biol.* **26**, 581–595 (1994).
- Savidge, B., Rounsley, S. D. & Yanofsky, M. F. Temporal relationship between the transcription of two *Arabidopsis* MADS box genes and the floral organ identity genes. *Plant Cell* **7**, 721–733 (1995).
- Töröner, W. *et al.* *GLOBOSA*: a homeotic gene which interacts with *DEFICIENS* in the control of *Antirrhinum* floral organogenesis. *EMBO J.* **11**, 4693–4704 (1992).
- Fan, H.-Y., Hu, Y., Tudor, M. & Ma, H. Specific interactions between the K domains of *AG* and *AGLs*, members of the MADS domain family of DNA binding proteins. *Plant J.* **11**, 999–1010 (1997).
- Davies, B., Egea-Cortines, M., de Andrade Silva, E., Saedler, H. & Sommer, H. Multiple interactions amongst floral homeotic MADS box proteins. *EMBO J.* **15**, 4330–4343 (1996).
- Egea-Cortines, M., Saedler, H. & Sommer, H. Ternary complex formation between the MADS-box proteins *SQUAMOSA*, *DEFICIENS* and *GLOBOSA* is involved in the control of floral architecture in *Antirrhinum majus*. *EMBO J.* **18**, 5370–5379 (1999).

23. Ambrose, B. A. *et al.* Molecular and genetic analyses of the *Silky1* gene reveals conservation in floral organ specification between eudicots and monocots. *Mol. Cell* (in the press).
24. Angenent, G. C., Franken, J., Busscher, M., Weiss, D. & van Tunen, A. J. Co-suppression of the petunia homeotic gene *flp2* affects the identity of the generative meristem. *Plant J.* 5, 33–44 (1994).
25. Pnueli, L., Hareven, D., Broday, L., Hurwitz, C. & Lifschitz, E. The *TM5* MADS box gene mediates organ differentiation in the three inner whorls of tomato flowers. *Plant Cell* 6, 175–186 (1994).
26. Kang, H.-G. & An, G. Isolation and characterization of a Rice MADS box gene belonging to the *AGL2* gene family. *Mol. Cell* 7, 45–51 (1997).
27. Mouradov, A. *et al.* Family of MADS-box genes expressed early in male and female reproductive structures of Monterey pine. *Plant Physiol.* 117, 55–62 (1998).
28. Mizukami, Y. & Ma, H. Ectopic expression of the floral homeotic gene *AGAMOUS* in transgenic *Arabidopsis* plants alters floral organ identity. *Cell* 71, 119–131 (1992).
29. Krizek, B. A. & Meyerowitz, E. M. The *Arabidopsis* homeotic genes *APETALA3* and *PISTILLATA* are sufficient to provide the B class organ identity function. *Development* 122, 11–22 (1996).
30. Wisman, E. *et al.* Knock-out mutants from an *En-1* mutagenized *Arabidopsis thaliana* population generate phenylpropanoid biosynthesis phenotypes. *Proc. Natl Acad. Sci. USA* 95, 12432–12437 (1998).

Acknowledgements

We thank J. Bowman for the *pi ag* double mutant photograph, K. Goto for sharing unpublished results, C. Wiley, C. Kawashima and T. Kawashima for technical assistance, D. Weigel, S. Liljegren, J. Bowman, D. Smyth and E. Meyerowitz for comments on the manuscript; and T. Casper and DuPont for access to pooled T-DNA insertion lines. S.P. was supported in part by the Human Frontiers Science Program Organization, and this work was supported by grants from the National Science Foundation and the National Institutes of Health (M.Y.).

Correspondence and requests for materials should be addressed to M.Y. (e-mail: marty@ucsd.edu).

Disruption of the plant gene *MOM* releases transcriptional silencing of methylated genes

Paolo Amedeo*, Yoshiki Habu*, Karin Afsar, Ortrun Mittelsten Scheid & Jerzy Paszkowski

Friedrich Miescher Institute, PO Box 2543, CH-4002 Basel, Switzerland
* These authors contributed equally to this work.

Epigenetic modifications change transcription patterns in multicellular organisms to achieve tissue-specific gene expression and inactivate alien DNA such as transposons or transgenes^{1,2}. In plants and animals, DNA methylation is involved in heritability and flexibility of epigenetic states³, although its function is far from clear. We have isolated an *Arabidopsis* gene, *MOM*, whose product is required for the maintenance of transcriptional gene silencing. Mutation of this gene or depletion of its transcript by expression of antisense RNA reactivates transcription from several previously silent, heavily methylated loci. Despite this, the dense methylation at these reactivated loci is maintained even after nine generations, indicating that transcriptional activity and methylation pattern are inherited independently. The predicted *MOM* gene product is a nuclear protein of 2,001 amino acids containing a region similar to part of the ATPase region of the SWI2/SNF2 family, members of which are involved in chromatin remodelling⁴. *MOM* is the first known molecular component that is essential for transcriptional gene silencing and does not affect methylation pattern. Thus, it may act downstream of methylation in epigenetic regulation, or be part of a new pathway that does not require methylation marks.

Cytosine residues in promoters of transcriptionally silent genes are frequently hypermethylated. Methylation patterns can be clonally transmitted by a maintenance system coupled to DNA replication, but they can also be modified by *de novo* methylation⁵ or loss of methylcytosine⁶. In *Arabidopsis*, deficiencies in methyltransferase^{7,8} or mutations in a SWI2/SNF2-like protein reduce methylation and

interfere with the maintenance of transcriptional gene silencing (TGS)^{9,10}. These observations support the hypothesis that methylation changes are a prerequisite for epigenetic shifts of transcriptional activity³. Methylation-mediated TGS may involve proteins with methylcytosine-binding domains (MBD), such as MeCP2, and their association with histone deacetylase complexes to produce transcriptionally repressive chromatin^{11,12}.

To identify molecular elements of TGS, we used random insertion of *Agrobacterium* transferred DNA (T-DNA) to mutate a previously characterized *Arabidopsis* line. This line is homozygous for a transgenic *HPT* locus that encodes hygromycin resistance but is hypermethylated and transcriptionally silenced (line A; ref. 10 and data not shown). We screened the progeny of 3,985 independent transformants (M2) for hygromycin resistance to identify mutations interfering with the maintenance of TGS and reactivating the silent locus. One M2 family segregated approximately 25% hygromycin-resistant seedlings. RNA blot hybridization showed that the transcriptional activity of the *HPT* gene was restored (Fig. 1a). To examine the activity of the mutation towards other transcriptionally silenced target loci, we crossed the mutant with a transgenic strain containing a β -glucuronidase (GUS) transgene silenced by TGS (line 6b5; H. Vaucheret, personal communication).

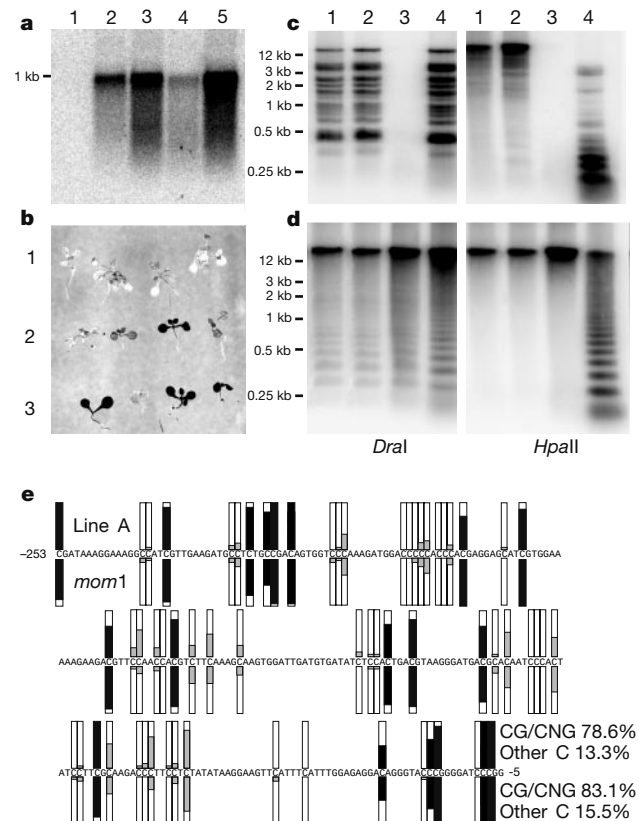


Figure 1 Analysis of *HPT* transcripts and DNA methylation status in the *mom1* mutant. **a**, Northern blot of RNA from the silent Line A (lane 1) and four M2 family *mom1* siblings selected on hygromycin (lanes 2–5), hybridized to *HPT*. **b**, Histochemical GUS staining of 6b5 seedlings (row 1), non-selected F2 seedlings from a cross between 6b5 and *mom1* (row 2) and F2 seedlings pre-selected on hygromycin for the *mom1/mom1* genotype (row 3). **c**, Southern blot hybridization of DNA from line A (lane 1), *mom1* (lane 2), wild type (lane 3) and som5 (lane 4; ref. 10), restricted with methylation-insensitive (*DraI*, left) and methylation-sensitive (*HpaII*, right) enzymes, and hybridized with the *HPT* gene linked to the 35S promoter. **d**, The same blot hybridized with a 180-bp pericentromeric repeat¹⁴. **e**, Genomic sequencing of the 35S promoter after bisulphite treatment in line A (above sequence) and in the *mom1* mutant (below sequence). The percentage of methylation at a particular site is indicated in black (methylation at CG or CNG sites) and grey (methylation at other sites).

Expression was reactivated in 18% of seedlings from the unselected segregating F2 and in 75% of F2 plantlets pre-selected for a mutant genotype by their hygromycin resistance (Fig. 1b). These data indicate the presence of a transacting recessive mutation affecting TGS in general.

Southern blot analysis using the methylation-sensitive restriction enzymes *HpaII* (CG methylation) and *MspI* (CNG methylation) indicated that the transgenic loci, although active, remained hypermethylated (Fig. 1c and data not shown). This resembles the silencing mutants *sil1* and *sil2* (ref. 13). The genes affected in *sil1* and *sil2* have not been characterized, but complementation tests revealed that they are not allelic to our new mutation (data not shown). We named the new silencing mutant *mom1* for mutation in a 'Morpheus' molecule.

Although the *HPT* gene was continuously expressed, hypermethylation of the locus was maintained at the level of the silent progenitor line for at least nine generations. Moreover, there was no change in the methylation of the 180-base-pair (bp) pericentromeric repeats (Fig. 1d), indicating unchanged levels of methylation throughout the genome¹⁴. A more refined analysis of the *HPT* promoter region by genomic sequencing after treatment with bisulphite also showed no significant changes in the distribution of methylation (Fig. 1e). Although a subtle effect of *mom1* on methylation cannot be excluded, these findings contrast with previously characterized mutations (*ddm1* for decreased DNA methylation, and alleles of *ddm1/som4–8*) that counteract TGS by a substantial reduction of DNA methylation levels^{10,15}.

A peculiarity of the *ddm1/som* mutants is the accumulation of developmental abnormalities. Inbreeding of these mutants for three or four generations resulted in weak, highly aberrant plants (Fig. 2a), perhaps reflecting ectopic, unscheduled expression or repression of genes involved in development¹⁶. Although the *ddm1/som* mutations are recessive, their outcrosses showed very slow resiliencing of the target genes, requiring the presence of the wild-type gene product for several generations to regain full silencing¹⁶. In contrast, *mom1* plants were indistinguishable from the progenitor Line A, even after inbreeding for nine generations (Fig. 2a). Resiliencing of loci reactivated by *mom1* occurred immediately after introduction of a wild-type *MOM* allele in F1 hybrids (Fig. 2b).

To identify the *MOM* gene, we subjected the DNA of a plant homozygous for a T-DNA insert that co-segregated with the mutant phenotype to thermal asymmetric interlaced polymerase chain reaction (TAIL-PCR) and used the amplified fragment to screen a genomic library of wild-type *Arabidopsis*. Multiple screening and

sequencing allowed assembly of a chromosomal sequence of 14 kilobases (kb). This sequence is located at the top of chromosome I, close to the *phyA* locus, and it is partially present on bacterial artificial chromosome (BAC) F3G11. Comparison of the wild-type with the mutant locus revealed a deletion of 1,980 bp that had been replaced by the T-DNA insertion (Fig. 3a). On northern blots, we hybridized probes flanking both sides of the deletion with *mom1* and wild-type RNA (Fig. 3a). The probes detected two different transcripts, both of which were affected in the *mom1* mutant. The wild-type 2.3-kb transcript was replaced in the mutant by two longer (2.7 and 2.4 kb) and more abundant transcripts. The 6.8-kb transcript was replaced in *mom1* by two new transcripts of 6.6 kb and 7.0 kb (Fig. 3a), probably resulting from two alternative polyadenylation sites encoded by the T-DNA. Thus, the T-DNA insertion affected expression of two neighbouring genes. We isolated full-length complementary DNAs for both genes by a combination of cDNA library screening, PCR after reverse transcription of RNA (RT-PCR), and 5' rapid amplification of cloned ends (RACE).

To determine which of the two genes is *MOM*, we generated transgenic plants expressing RNA antisense to each transcript. The empty vector construct (five lines) and antisense expression of the 2.3-kb cDNA (20 lines) did not interfere with TGS (Fig. 3b and data not shown). However, constructs containing the 3' part of the 6.8-kb cDNA in an antisense orientation reactivated the silent transgenic locus in 15 out of 20 independent transformants (Fig. 3b). Most important, this alleviation of TGS also occurred without a change in methylation levels (data not shown), perfectly recreating the *mom1* mutant phenotype and demonstrating that this is the cDNA for *MOM*.

The *MOM* gene is present in a single copy in the *Arabidopsis* genome. It is composed of 16 exons and encodes a protein of 2,001 amino acids containing three types of internal repeats (Fig. 4a). The *MOM* protein has no overall homology to sequences deposited in databases, but contains two regions with similarities to known proteins. The most extended sequence similarity is to the ATPase region of the SWI2/SNF2 family⁴ which is involved in chromatin remodelling (Fig. 4b). Proteins of this family contain sequence motifs similar to those of DNA and RNA helicases belonging to superfamily II (ref. 17). According to their crystal structure, helicases have two distinct domains separated by a deep cleft^{18,19}.

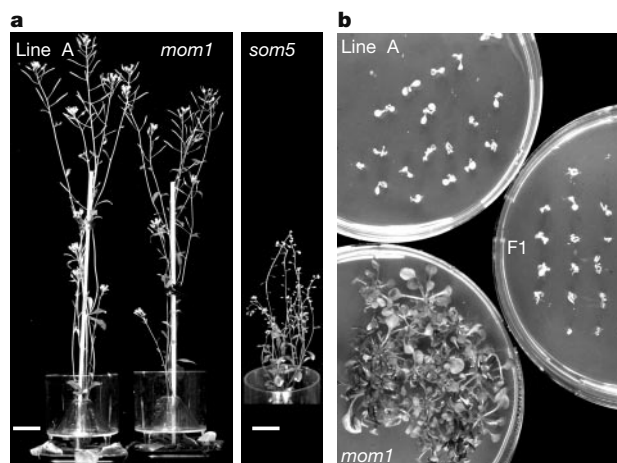


Figure 2 Phenotypes of the *mom1* mutant. **a**, Line A, *mom1* and *som5* after selfing for six generations (white bar 2.5 cm). **b**, Immediate resiliencing of the *HPT* transgene in a backcross of *mom1* to line A. Seeds of line A, *mom1* and the F1 hybrid were germinated on a medium containing hygromycin.

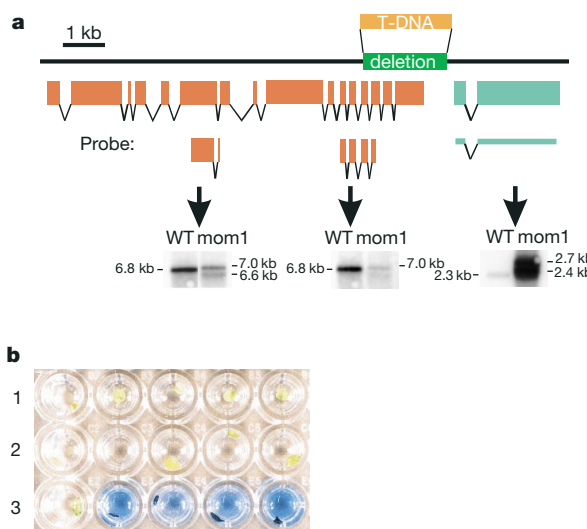


Figure 3 Structure and expression of the mutant *mom1* locus and identification of the *MOM* gene. **a**, The T-DNA insert affected transcripts of two neighbouring genes, as shown on the northern blots below. **b**, Reactivation of the silenced *GUS* transgene in leaves of 6b5 plants expressing part of *MOM* mRNA in the antisense orientation. Row 1, 6b5 plants; row 2, 6b5 plants transformed with the empty vector; row 3, 6b5 plants transformed with a vector designed for antisense inhibition of the *MOM* gene.

Structural analysis of helicase–DNA complexes indicates functional cooperation and partial independence of the two domains²⁰. The region present in MOM spans only the second domain, thus representing half of a helicase unit. An attractive possibility is that MOM functions as a heterodimer with a still unknown partner carrying the missing half of the helicase region. Protein sequences containing complementary halves of helicases are known for bacteria²¹.

Actin and actin-related proteins present in the nucleus are involved in epigenetic regulation and chromatin remodelling^{4,22}. A

stretch of 100 amino acids close to the carboxyl-terminal end of MOM has a similarity to a short region of chicken tensin (Fig. 4c). This particular tensin fragment contains direct repeats and acts as a barbed-end actin-binding domain²³. In addition, MOM has three predicted nuclear localization signal (NLS) sequences²⁴ and did in fact localize to the nucleus when overexpressed (Fig. 4d). It also has a putative transmembrane domain (TD) (Fig. 4a) predicted by hydrophobicity analysis and the positive-inside rule²⁵. This arrangement (NLS and TD) resembles integral membrane proteins of the inner nuclear membrane, such as the lamin B receptor²⁶ or the

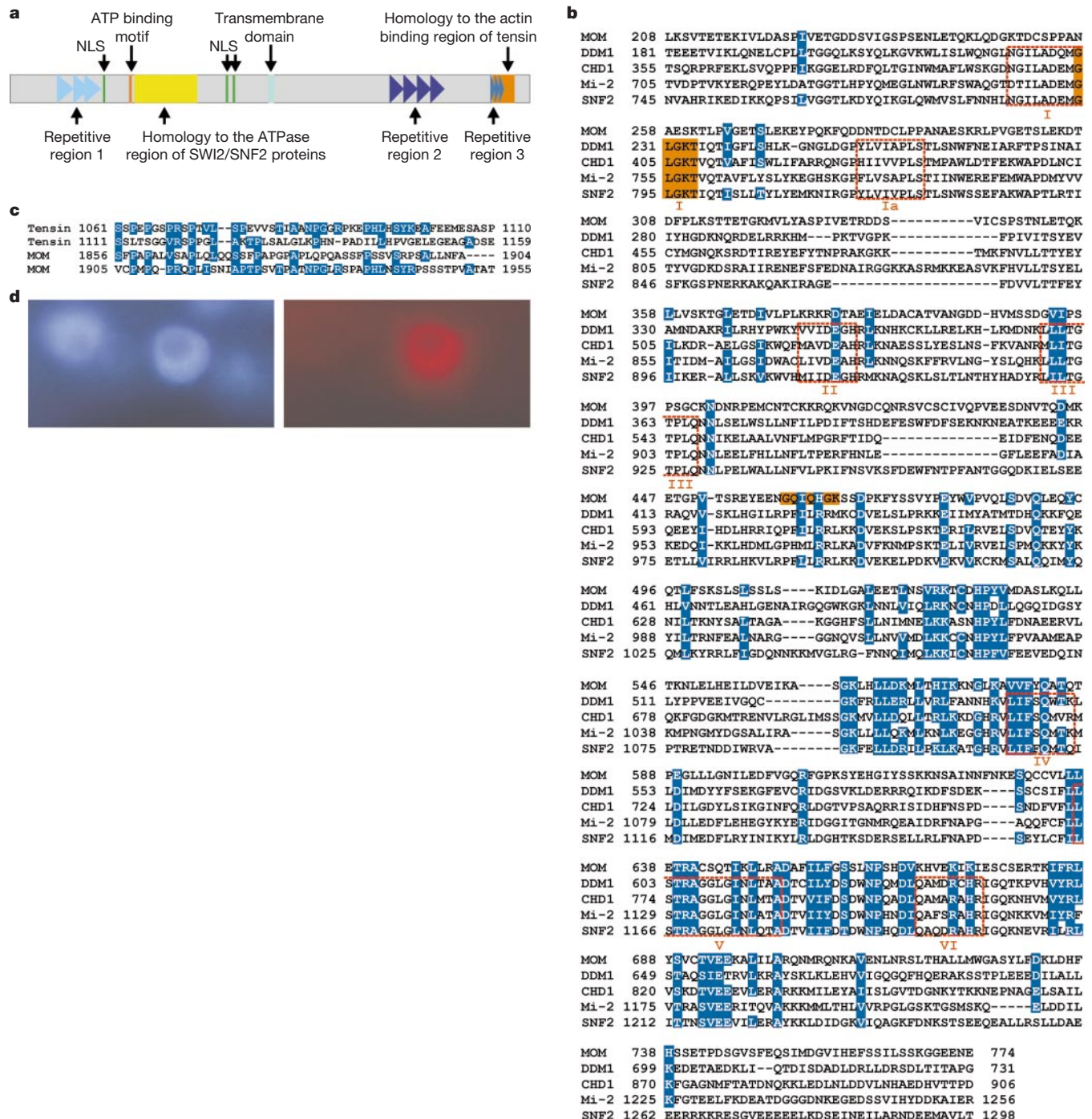


Figure 4 Comparison of MOM with other proteins and its subcellular localization. **a**, Distribution of MOM domains and regions of similarity to other proteins. **b**, Comparison of the amino-acid sequences of MOM with other ATPases in the SWI2/SNF2 family: *Arabidopsis* DDM1, yeast CHD1, human Mi-2 and yeast SNF2. The helicase motifs are boxed in red and predicted ATP/GTP-binding sites¹⁷ are coloured red. Identical amino-

acid residues or those replaced by conservative substitutions in MOM and present in at least three of the other proteins are coloured blue. **c**, Comparison of the amino-acid repeats of MOM and the actin-binding region of chicken tensin. Conserved amino-acid residues are coloured blue. **d**, Nuclear localization of MOM. Left, DAPI staining of nuclei; right, subcellular immunodetection of HA-epitope-tagged MOM.

lamin-associated proteins LAP1 and LAP2 (refs 27, 28). Thus, MOM may localize to the inner membrane of the nucleus and interact with actin or actin-related proteins, in addition to contacting DNA through the 'half helicase' region.

Two alternative functions of MOM in TGS can be envisaged; either MOM is required for recognition of methylation as a silencing signal and thus acts downstream of methylation, or MOM acts independently of methylation. The failure of overall methylation recognition in the absence of MOM should result in a phenotype similar to that of mutants affected in DNA methylation itself, such as *ddm/som*. However, the normal growth of *mom1* clearly differs from *ddm/som*, indicating that the action of MOM in the silencing process is more restricted and possibly confined to selected chromosomal information. The specificity of methylation signals may be achieved by the diversity of MBD proteins and their ability to participate in various protein complexes affecting chromatin structure²⁹. Thus, if MOM is involved in methylation-mediated TGS, it could be a part of such a complex with selected silencing targets.

Alternatively, MOM may be involved in the regulation of transcriptional activity by a mechanism independent of a DNA methylation signal. If this is the case, it suggests two parallel and independent levels of epigenetic regulation in organisms exploiting DNA methylation. In either case, MOM is the first well defined element able to modify an epigenetic state in a methylation-independent fashion. □

Methods

Mutagenesis and selection

T-DNA transformation was achieved by *in planta* vacuum infiltration. We pooled selfed seeds from individual transformants in aliquots of approximately 50 seeds each from 20 T2 families and screened for hygromycin resistance on agar-solidified germination medium containing 10 mg l⁻¹ hygromycin B (ref. 10).

Northern and Southern blots and methylation analysis

We prepared and hybridized genomic DNA and RNA as described previously¹⁰. DNA methylation was analysed either by Southern blotting using methylation-sensitive restriction enzymes or by genomic sequencing after bisulphite treatment. We performed genomic sequencing of the 35S promoter core region in line A and in *mom1* with primers designed for a preferential amplification of unmethylated templates. We sequenced 22 individual clones originating from four independent PCR reactions for each genotype.

Isolation of the MOM gene

After genetically separating the two T-DNA inserts present in the M2 family, we cloned a DNA sequence flanking the co-segregating T-DNA insert by TAIL-PCR using three specific, nested primers close to the right border of the T-DNA and directed outwards (5'-CATCTACGCAATGTACCAGC-3'; 5'-GATGGGAATGGCTGAGTGGC-3'; 5'-CAGTTCACCAACGTAAACGGC-3') in combination with two degenerate primers (5'-WGCNAGTNAGWANAA G-3' or 5'-AWGCANGNCWGANATA-3'). The larger fragment (325 bp) was then cloned and used to probe Southern blots and to screen a genomic library (Stratagene) and a cDNA library. The genomic region around the T-DNA was assembled from different overlapping clones. We cloned the partial cDNAs of the two genes flanking the T-DNA from a cDNA library, RT-PCR and 5'RACE products. The cDNA fragments used for the antisense approach were cloned into pbarbi53, a modified version of p1'barbi (ref. 30). This carries an expression cassette consisting of the 35S promoter of cauliflower mosaic virus (CaMV), a multiple cloning site and the 35S terminator of CaMV. We used the resulting recombinant plasmids to transform the transgenic plant line 6b5 of *Arabidopsis thaliana* containing transcriptionally silenced multiple copies of the GUS gene (obtained from H. Vaucheret). The antisense transformants and empty vector controls were examined for reactivation of the GUS gene by histochemical staining.

Sequence analysis

We performed sequence comparisons using GCG (Genetic Computer Group) and DNASIS (Hitachi). Parts of the genomic sequence and of the cDNA were found in the *Arabidopsis* database (accession numbers B67281, B62563, B20434, B20425, B21274, B08967, B11993, B20116, B12496, B10852, Z18494, AA597930, AL081635, AL081579, AL091323, AL082308 and B12172).

Subcellular localization of MOM

We inserted a DNA fragment encoding a HA-tag generated by annealing two oligonucleotides (5'-GATCTATATCCATATGATGTTCTCTGATTATGCTTGA-3'; 5'-GATCTCAAGCATAATCAGGAACATCATATGGATATA-3') in frame 3' of the MOM open reading frame. After inserting the resulting cDNA fragment into a plant expression vector (pCAMBIA2200) behind the CaMV 35S promoter, we introduced this construct into

tobacco BY-2 cells by particle bombardment. After 24-h incubation, cells were fixed with paraformaldehyde, placed on a polylysine-coated glass slide and treated briefly with a mixture of 0.002% pectolyase, 0.01% macerozyme, 0.03% catylase. For intracellular detection of MOM, we stained cells with anti-HA-epitope antibodies and anti-mouse immunoglobulin antibodies conjugated with Texas Red and counter-stained nuclei with 4,6-diamidino-2-phenylindole (DAPI).

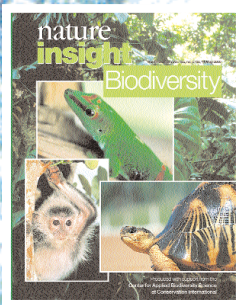
Received 16 December 1999; accepted 10 March 2000.

- Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genet.* **20**, 116–118 (1998).
- Matzke, A. J. M. & Matzke, M. A. Position effects and epigenetic silencing of plant transgenes. *Curr. Opin. Plant Biol.* **1**, 142–148 (1998).
- Bird, A. P. Gene number, noise reduction and biological complexity. *Trends Genet.* **11**, 94–100 (1995).
- Kingston, R. E. & Narlikar, G. J. ATP-dependent remodelling and acetylation as regulators of chromatin fluidity. *Genes Dev.* **13**, 2339–2352 (1999).
- Adams, R. L. P. et al. in *DNA Methylation: Molecular Biology and Biological Significance* (eds Jost, J. P. & Saluz, H. P.) 120–144 (Birkhauser, Basel, Boston, Berlin, 1993).
- Jost, J. P., Siegmund, M., Sun, L. & Leung, R. Mechanism of DNA demethylation in chicken embryos: Purification and properties of a 5-methylcytosine-DNA glycosidase. *J. Biol. Chem.* **270**, 9734–9739 (1995).
- Ronemus, M. J., Galbiati, M., Ticknor, C., Chen, J. & Dellaporta, S. L. Demethylation-induced developmental pleiotropy in *Arabidopsis*. *Science* **273**, 654–656 (1996).
- Finnegan, E. J., Peacock, W. J. & Dennis, E. S. Reduced DNA methylation in *Arabidopsis thaliana* results in abnormal plant development. *Proc. Natl Acad. Sci. USA* **93**, 8449–8454 (1996).
- Jeddeloh, J. A., Bender, J. & Richards, E. J. The DNA methylation locus DDM1 is required for maintenance of gene silencing in *Arabidopsis*. *Genes Dev.* **12**, 1714–1725 (1998).
- Mittelman, S. C., Olszewski, K. & Paszkowski, J. Release of epigenetic gene silencing by trans-acting mutations in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **95**, 632–637 (1998).
- Nan, X. S. et al. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386–389 (1998).
- Jones, P. L. et al. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genet.* **19**, 187–191 (1998).
- Furner, I. J., Sheikh, M. A. & Collett, C. E. Gene silencing and homology-dependent gene silencing in *Arabidopsis*: genetic modifiers and DNA methylation. *Genetics* **149**, 651–662 (1998).
- Vongs, A., Kakutani, T., Martienssen, R. A. & Richards, E. J. *Arabidopsis thaliana* DNA methylation mutants. *Science* **260**, 1926–1928 (1993).
- Jeddeloh, J. A., Stokes, T. L. & Richards, E. J. Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nature Genet.* **22**, 94–97 (1999).
- Kakutani, T., Munakata, K., Richards, E. J. & Hirochika, H. Meiotically and mitotically stable inheritance of DNA hypomethylation induced by *ddm1* mutation of *Arabidopsis thaliana*. *Genetics* **151**, 831–838 (1999).
- Gorbalenya, A. E. & Koonin, E. V. Helicases: amino acid sequence comparisons and structure–function relationships. *Curr. Opin. Struct. Biol.* **3**, 419–429 (1993).
- Subramanya, H. S., Bird, L. E., Brannigan, J. A. & Wigley, D. B. Crystal structure of a Dext box DNA helicase. *Nature* **384**, 379–383 (1996).
- Yao, N. et al. Structure of the hepatitis C virus RNA helicase domain. *Nature Struct. Biol.* **4**, 463–497 (1997).
- Velankar, S. S., Soultanas, P., Dillingham, M. S., Subramanya, H. S. & Wigley, D. B. Crystal structures of complexes of PcrA DNA helicases with a DNA substrate indicate an inchworm mechanism. *Cell* **97**, 75–84 (1999).
- Koonin, E. V. & Rudd, K. E. Two domains of superfamily I helicases may exist as separate proteins. *Protein Sci.* **5**, 178–180 (1996).
- Jiang, Y. W. & Stillman, D. Epigenetic effects on yeast transcription caused by mutations in an actin-related protein present in the nucleus. *Genes Dev.* **10**, 604–619 (1996).
- Chuang, J. Z., Lin, D. C. & Lin, S. Molecular cloning, expression, and mapping of the high affinity actin-binding domain of chicken cardiac tensin. *J. Cell Biol.* **128**, 1095–1109 (1995).
- Hicks, G. R. & Raikhel, N. V. Protein import into the nucleus: an integrated view. *Annu. Rev. Cell Dev. Biol.* **11**, 155–188 (1995).
- von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487–494 (1992).
- Workman, H. J., Evance, C. D. & Blobel, G. The lamin B receptor of the nuclear envelope inner membrane: a polytopic protein with eight potential transmembrane domains. *J. Cell Biol.* **111**, 1535–1542 (1990).
- Martin, L., Crimando, C. & Gerace, L. cDNA cloning and characterization of lamina-associated polypeptide 1C (LAP1C), an integral protein of the inner nuclear membrane. *J. Biol. Chem.* **270**, 8822–8828 (1995).
- Furukawa, K., Pante, N., Aebi, U. & Gerace, L. Cloning of a cDNA for lamina-associated polypeptide 2 (LAP2) and identification of regions that specify targeting to the nuclear envelope. *EMBO J.* **14**, 1626–1636 (1995).
- Bird, A. & Wolffe, A. P. Methylation-induced repression—belts, braces, and chromatin. *Cell* **99**, 451–454 (1999).
- Mengiste, T., Amedeo, P. & Paszkowski, J. High-efficiency transformation of *Arabidopsis thaliana* with a selectable marker gene regulated by the T-DNA 1' promoter. *Plant J.* **12**, 945–948 (1997).

Acknowledgements

We thank H. Vaucheret for the *Arabidopsis* line 6b5, M. Duckely for help with nuclear localization of MOM, and M. Briker and S. van Eeden for greenhouse help. We also thank B. Hohn, J.-P. Jost, F. Meins, J. Hofsteenge and P. King for their comments on the manuscript. Y. Habu is a visiting scientist from the National Institute for Basic Biology, Okazaki, Japan.

Correspondence and requests for materials should be addressed to Y. Habu (e-mail: habu@fmi.ch). Accession numbers for the complete MOM genomic sequence (Co) and cDNA (Zh) are AF213628 and AF213627, respectively.



Cover illustrations
courtesy of
Conservation
International.

nature insight

Reprinted from Vol. 405, no. 6783, 11 May 2000

Biodiversity

Biodiversity may be a buzzword, but as a concept it sits at the heart of ecological research. Some ecological communities, such as pristine coral reef systems, are astonishingly rich in the number and types of species that they support, whereas others are relatively species poor. Natural communities also differ greatly in the proportion of species performing different ecological functions. What determines such differences and how these differences are related to ecosystem functioning are questions that have occupied the minds of ecologists for decades.

But these questions are so much more pressing now. We live at a time of rapid environmental change, resulting largely from our own activities, and a concomitant, accelerating rate of habitat loss and species extinctions. Like children playing with fire, we do not fully understand, and therefore cannot predict, the ultimate consequences of tampering with global biodiversity. This collection of reviews — the second in our new section called 'Nature Insight' — focuses on the science of biodiversity.

We are pleased to acknowledge the financial support of the Center for Applied Biodiversity Science (CABS), a division of Conservation International, in producing this Insight. The content is in accord with the philosophy that biodiversity conservation is a human-centred pursuit that must be underpinned by solid science. Of course, *Nature* carries sole responsibility for all editorial content and rigorous peer-review.

This Insight is deliberately broad in scope, covering underlying concepts, pure and applied research, and biodiversity loss from the human perspective. We hope that scientists, policy-makers and general readers alike will find the reviews both informative and thought provoking. Given that environmental change and biodiversity loss is a global concern, and understanding that not everyone will have easy access to the print version, this Insight is freely available to all readers, regardless of subscriber status, on our website at www.nature.com.

Nature London
Porters South, 4 Crinan St,
London N1 9XW, UK
Tel +44 207 833 4000
Fax +44 207 843 4596/7
e-mail: nature@nature.com
<http://www.nature.com>

Nature Washington
968 National Press Building,
529 14th St NW,
Washington DC 20045, USA
Tel +1 202 737 2355
Fax +1 202 628 1609
e-mail: nature@nature.com
<http://www.nature.com>

Nature Tokyo
Shin-Mitsuke Building (4F),
3-6 Ichigaya Tamachi,
Shinjuku-ku,
Tokyo 162, Japan
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: nature@naturejpn.com
<http://www.naturejpn.com>



Editor, Nature: Philip Campbell
Insight Editors: Ritu Dhand,
Rory Howlett

Production Editor: Simon Gribbin
Art Director: Majo Xeridat

Layout: Nicola Barker,
Suzanne Coleman,
Fiona Rawlinson

Production Manager: Yvonne Strong
Publisher: Liz Allen

overview:

- 208 Causes, consequences and ethics of biodiversity**
D. Tilman

review articles:

- 212 Getting the measure of biodiversity**
A. Purvis & A. Hector

- 220 Global patterns in biodiversity**
K. J. Gaston

- 228 The diversity–stability debate**
K. S. McCann

- 234 Consequences of changing biodiversity**
F. S. Chapin III, E. S. Zavaleta, V. T. Eviner, R. L. Naylor, P. M. Vitousek, H. L. Reynolds, D. U. Hooper, S. Lavorel, O. E. Sala, S. E. Hobbie, M. C. Mack & S. Díaz

- 243 Systematic conservation planning**
C. R. Margules & R. L. Pressey

Nature (ISSN 0028-0836) is published weekly on Thursday, except the last week in December, by Nature Publishing Group (Porters South, 4 Crinan Street, London N1 9XW). Registered as a newspaper at the British Post Office. Annual subscription for the Americas US\$595 (institutional/corporate), US\$159 (individual making personal payment). Canada residents please add 7% GST (No. 140911595). North and South American orders to: *Nature*, Subscription Dept, P. O. Box 5055, Brentwood, TN 37024-5055, USA. Other orders to *Nature*, Brunel Road, Basingstoke, Hants RG21 2XS, UK. Periodicals postage paid at New York, NY 10010-1707, and additional mailing offices. Authorization to photocopy material for internal or personal use, or internal or personal use of specific clients, is granted by *Nature* to libraries and others registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided the base fee of \$12.00 an article (or \$2.00 a page) is paid direct to CCC, 222 Rosewood Drive, Danvers, MA 01923, USA. Identification code for *Nature*: 0028-0836/00 \$12.00+\$2.00. US Postmaster send address changes to: *Nature*, PO Box 5055, Brentwood, TN 37024-5055. Published in Japan by Nature Japan K.K., Shin-Mitsuke Bldg. 36 Ichigaya Tamachi, Shinjuku-ku, Tokyo 162, Japan. © 2000 Nature Publishing Group.

Causes, consequences and ethics of biodiversity

David Tilman

Department of Ecology, Evolution and Behavior, University of Minnesota, St Paul, Minnesota 55108, USA (e-mail: tilman@lter.umn.edu)

The existence of so great a diversity of species on Earth remains a mystery, the solution to which may also explain why and how biodiversity influences the functioning of ecosystems. The answer may lie in quantifying the trade-offs that organisms face in dealing with the constraints of their environment. Societal responses to the loss of biodiversity also involve trade-offs, and the elaboration of these will be essential in developing wiser environmental ethics and policy.

The most striking feature of Earth is the existence of life, and the most striking feature of life is its diversity. This biological diversity, or biodiversity, has long been a source of wonderment and scientific curiosity, but is increasingly a source of concern. Human domination of Earth's ecosystems¹ is markedly reducing the diversity of species within many habitats worldwide, and is accelerating extinction. One of the more pragmatic questions raised by these threats to biodiversity is the extent to which this loss of biodiversity matters; that is, are stability, productivity and other aspects of the functioning of both managed and natural ecosystems dependent on biodiversity?

There are strong reasons to hypothesize, as did Darwin² and Elton³, that biodiversity might impact ecosystem processes. But ecology is no longer a discipline in which natural history observations and simple verbal logic hold sway. The rekindled interest in the potential effects of biodiversity on ecosystem processes, which followed the publication in 1993 of a book edited by Schulze and Mooney⁴, is occurring in a discipline for which hypotheses are now tested against the results of field experiments, mechanistic theory and quantitative field observations. Anything less than the concordance of all three lines of evidence leads to the modification or rejection of hypotheses. Given that this topic became a principal focus of scientific inquiry only about seven years ago, it is not surprising that it remains contentious. Indeed, the greatest surprise may be the rapidity, breadth and depth of work that already has occurred, and the generalities that are emerging from it.

Five papers in this issue summarize this work. Purvis and Hector (pages 212–219), McCann (pages 228–233), and Chapin and collaborators (pages 234–242) review and synthesize recent experimental, theoretical and observational studies that have demonstrated links between biodiversity and the stability, productivity and nutrient dynamics of ecosystems. Gaston (pages 220–227) summarizes global patterns of biodiversity and some possible explanations for these patterns. Margules and

Pressey (pages 243–253) discuss strategies for the preservation of biodiversity.

The effects of biodiversity on ecosystems

In broad summary, these reviews show that, on average, greater diversity leads to greater productivity in plant communities, greater nutrient retention in ecosystems and greater ecosystem stability. For instance, grassland field experiments both in North America (Fig. 1)^{5,6} and across eight different European sites, ranging from Greece in the south and east to Portugal and Ireland in the west and Sweden in the north⁷, have shown that each halving of the number of plant species within a plot leads to a 10–20% loss of productivity. An average plot containing one plant species is less than half as productive as an average plot containing 24–32 species^{5–7}. Lower plant diversity also leads to greater rates of loss of limiting soil nutrients through leaching, which ultimately should decrease soil fertility, further lowering plant productivity.

Both laboratory and field studies have shown that ecosystem processes are more variable (less stable or reliable) at lower diversity (see review by McCann, pages 228–233, and refs 8–10). The greater stability of more diverse ecosystems seems to result from three processes^{11–14}. The first is comparable to the economic process that causes a more diverse investment portfolio to be less volatile. Because species, like corporations, differ from each other,

they tend to respond somewhat independently to environmental variability. The more species

that such variability is averaged across, the less variable is their total¹¹. Second,

species within a given trophic level often compete with each other, which causes their abundances to negatively covary. When one species declines, another is freed

from competition and increases. This negative

covariance reduces the variability of the community as a whole^{13,14}. Finally, measures

of temporal stability compute variability relative to mean abundance, such as by using the ratio of community abundance to its temporal standard deviation. The tendency for community

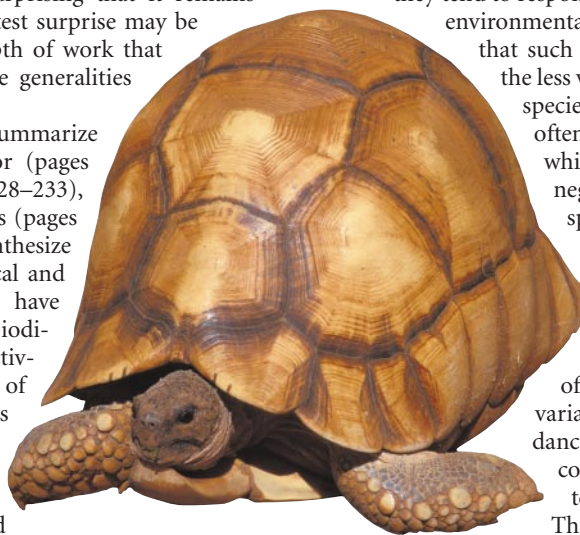




Figure 1 Biodiversity experiments, such as this one in Minnesota⁶ or the other experiments reviewed by Chapin *et al.* (pages 234–242) and by Purvis and Hector (pages 212–219), have shown that a greater number of plant species leads to greater community productivity. In the experiment shown, 245 plots, each 9 m × 9 m, were assigned randomly to have from 1 to 16 prairie plant species, with the species composition of each plot being separately chosen at random⁶. Species composition and plant diversity were both strong determinants of ecosystem functioning.

abundance to increase as diversity increases thus causes this ratio, which is a measure of stability, to increase as diversity increases¹⁴.

In total, biodiversity, which ten years ago was considered unimportant by most ecosystem ecologists, has now been shown to impact significantly upon many aspects of ecosystem functioning. Diversity must now be added to the list of factors — including species composition, disturbance regime, soil type and climate — that influence ecosystem functioning. The recent rediscovery of the importance of biodiversity highlights an under-appreciated truth — although society is dependent on natural and managed ecosystems for goods and services that are essential for human survival, we know all too little about how ecosystems work.

Two sets of unanswered scientific questions come to the forefront. First, why is the world so diverse; that is, what forces and processes led to the evolution and persistence of so many species? This is not merely an academic question. The processes that allow interacting species to coexist in an ecosystem simultaneously influence the productivity, nutrient dynamics and stability of that ecosystem. Second, what are the mechanisms by which the loss of diversity impacts the functioning of ecosystems, how general are these mechanisms, and how important is biodiversity relative to other factors that influence ecosystem functioning? In addition, the realization that human actions are harming, perhaps irreversibly, the ecosystems upon which humans depend raises a third, philosophical question: what should be the role of scientists and science in the development of ethics and policy?

Coexistence and ecosystem functioning

Both our understanding of the effects of biodiversity on ecosystem processes, and the effectiveness of alternative strategies for the preservation of biodiversity, are limited by our knowledge of the mechanisms that maintain diversity. The mechanisms most relevant to ecosystem functioning are those that maintain diversity on the

local scales within which individuals of one species interact with individuals of other species. It is from such interactions among individuals of different species that diversity is expected to impact ecosystem processes.

What are these mechanisms of coexistence? At present there are an abundance of alternative hypotheses but no clear demonstrations of the actual processes that maintain the diversity of species-rich ecosystems. In a general sense, coexistence requires the existence of evolutionarily persistent interspecific trade-offs in the abilities of species to deal with the factors that constrain their fitness and abundance. However, there are many potential constraints and trade-offs. Species may coexist because of interspecific trade-offs (1) between their competitive abilities and their dispersal abilities; (2) between their competitive abilities and their susceptibility to disease, herbivory or predation; (3) between their abilities to live off average conditions and their abilities to exploit resource pulses; or (4) between their abilities to compete for alternative resources in a heterogeneous landscape^{15–18}.

The effects on ecosystem functioning of many such mechanisms of coexistence have yet to be determined theoretically. However, it is already clear that the underlying mechanisms of coexistence can greatly influence how diversity affects ecosystem processes^{19,20}. Consider, for instance, plant species that coexist in a spatially heterogeneous habitat because of differences in both the soil pH and the temperature (which varies seasonally) at which each grows optimally (Fig. 2a). Such niche differentiation²⁰ causes the predicted productivity of plant communities to be an increasing function of plant diversity (Fig. 2b). Moreover, the pattern of this increase is such that there are some species combinations at a given level of diversity that are more productive than any possible combination of fewer species (Fig. 2b). The greater productivity of higher diversity communities occurs because, in such heterogeneous habitats, each species is a superior performer in only a portion of sites. Clearly, the magnitude

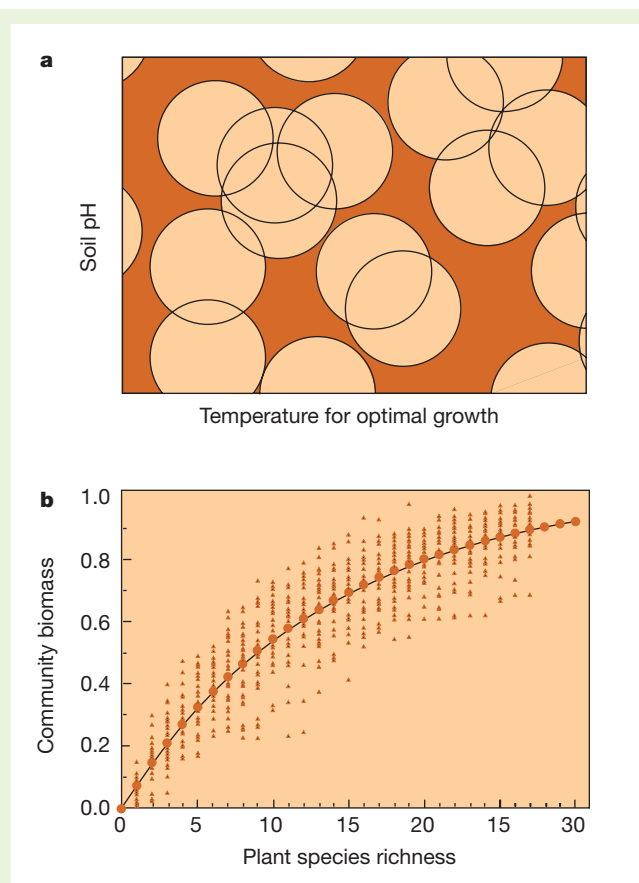


Figure 2 Niche differentiation and productivity. **a**, A simple model — the ‘snowballs on the barn’ model — of niche differentiation and coexistence²⁰. The range of conditions in which each species can exist is shown with a circle, the position of which is defined by its centre. By randomly choosing locations for various numbers of circles (species), it is possible to calculate the effect of diversity on the ‘coverage’ of the heterogeneous habitat. The amount of such coverage is proportional to community biomass. **b**, Results of simulations (triangles) and of an analytical solution (solid curve) to the effects of diversity on community productivity for the snowballs on the barn model²⁰.

of this effect increases as heterogeneity or diversity increase. Increased diversity leads, on average, to increased ‘coverage’ of the habitat conditions, that is, to increased efficiency of resource capture and use, because diversity increases the chance that the species that are better able to handle particular conditions are present. Assuming that species are chosen at random, diversity is a simple way to measure the range and coverage of species traits in a community.

In contrast, consider a case in which interspecific interactions are based on direct antagonism and not on efficiency of resource use. For a simple formulation, let there be an interspecific trade-off between competitive ability and productivity. Species that achieve greater productivity in monoculture would be poorer competitors, attaining lower abundances when competing. Because greater diversity increases the chance that a competitively superior but lower-yielding species would be present, productivity would, on average, be a decreasing function of diversity. This is a simple variant on the sampling-effect model^{20–22}, here modified to have better competitors be less, rather than more, productive.

What, then, is implied by available experimental results, which have shown that productivity is an increasing function of plant species diversity? They indicate that coexistence through niche differentiation and related processes may be more prevalent in nature than coexistence through antagonism and related processes, at least

for the types of communities studied so far. Expressed another way, much of nature may have a free-market economy, structured by the efficiencies of open competition among species, rather than an economy structured by pre-emption and other monopolistic practices. Such speculations may be premature, especially because complex systems containing many trophic levels (for example, plants, decomposers, herbivores and predators) are, as yet, poorly studied. However, they highlight the conceptual links between economics and ecology — disciplinary links that must be strengthened if ecological knowledge is to be used to help create a sustainable human economy.

Societal trade-offs and ethics

The progress made during the past seven years in understanding these issues underscores the potential implications of habitat simplification and loss of diversity for the ecosystem goods and services²³ upon which humans depend. The species presently inhabiting Earth are the result of over 3 billion years of natural selection that likely favoured efficiency, productivity and specialization. These organisms are the catalysts that capture and transform energy and materials, producing, among other things, food, fuel, fibre and medicines. These species recycle wastes, create pure drinking water, drive global biogeochemical cycles that created and maintain an aerobic atmosphere, regulate global climate through effects on greenhouse gases and local climate through effects on evapotranspiration, generate soil fertility, and provide other ecosystem goods and services²³. In addition, the Earth’s biodiversity is the source of all crops and all pollinators of crops, of all livestock, and of many pharmaceuticals and pesticides. Just three crops — corn, rice and wheat — provide about 60% of the human food supply. The viability of these crops depends on the maintenance of high genetic diversity²⁴, which can allow, among other things, development of strains that are resistant to emerging and evolving diseases and pests²⁵. In the long term, food stability will require development of new crops from what are now wild plants, because disease or pesticide-resistant pests will cause the loss of current crops, just as disease caused the loss of chestnut, elm and other tree species from North American forests.

Humans, like all other organisms, experience trade-offs. The loss of biodiversity will diminish the capacity of ecosystems to provide society with a stable and sustainable supply of essential goods and services, but many of the very actions that harm biodiversity simultaneously provide valuable societal benefits. There exists a trade-off defining the net benefits that society receives from the various ways that humans could use and impact nature, but, as yet, this is poorly defined. This trade-off itself is likely to shift through time in response to the remaining amounts and states of various resources, including biodiversity. The amounts and states of biotic resources have changed rapidly during the past century, as global population increased 3.7-fold and per capita gross domestic product, a reasonable proxy for consumption, increased 4.6-fold²⁶. It seems likely that environmental policy that is optimal from a societal perspective would be markedly different now from that of 250 years ago. However, we still use environmental and land-use ethics, codified in law, that were articulated during the era when the human population, at one-tenth its present size, tamed wilderness with axe and ox.

Science has much to contribute to dialogues on policy and ethics. Although academic institutions seem to value such contributions less than contributions to peer-reviewed journals, this is short-sighted. Ultimately, society invests in science because advances in scientific knowledge benefit society. The ethics of science cannot eschew involvement in public discourse. Science must contribute, in an open, unbiased manner, to relevant issues.

Because of the emergence of human domination of global ecosystems, society faces new, tough trade-offs. These include trade-offs between the current benefits and the future costs of environmental damage, and between benefits to a few and costs to many. Research is needed to quantify these trade-offs, and the work done so far on

biodiversity provides a good start. Additional work, at the interface between ecology and economics, is needed to quantify the immediate and long-term costs and benefits of alternative actions.

The world that will exist in 100 and 1,000 years will, unavoidably, be of human design, whether deliberate or haphazard. The principles that should guide this design must be based on science, much of it done only sketchily to date, and on ethics. Ethics should, among other things, apportion costs and benefits between individuals and society as a whole, and between current generations and all future generations. A sustainable world will require an ethic that is ultimately as incorporated into culture and as long lasting as a constitutional bill of rights or as religious commandments. The Earth will retain its most striking feature, its biodiversity, only if humans have the prescience to do so. This will occur, it seems, only if we realize the extent to which we use biodiversity. □

1. Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. Human domination of earth's ecosystems. *Science* **277**, 494–499 (1997).
2. Darwin, C. *The Origin of Species by Means of Natural Selection* (reprinted by The Modern Library, Random House, New York, 1859).
3. Elton, C. S. *The Ecology of Invasions by Animals and Plants* (Methuen, London, 1958).
4. Schulze, E. D. & Mooney, H. A. *Biodiversity and Ecosystem Function* (Springer, Berlin, 1993).
5. Tilman, D., Wedin, D. & Knops, J. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* **379**, 718–720 (1996).
6. Tilman, D., Knops, J., Wedin, D., Reich, P., Ritchie, M. & Siemann, E. The influence of functional diversity and composition on ecosystem processes. *Science* **277**, 1300–1302 (1997).
7. Hector, A. *et al.* Plant diversity and productivity experiments in European grasslands. *Science* **286**, 1123–1127 (1999).
8. Tilman, D. & Downing, J. A. Biodiversity and stability in grasslands. *Nature* **367**, 363–365 (1994).
9. McGrady-Steed, J., Harris, P. M. & Morin, P. J. Biodiversity regulates ecosystem predictability. *Nature* **390**, 162–165 (1997).
10. Naeem, S. & Li, S. Biodiversity enhances ecosystem reliability. *Nature* **390**, 507–509 (1997).
11. Doak, D. F. *et al.* The statistical inevitability of stability-diversity relationships in community ecology. *Am. Nat.* **151**, 264–276 (1998).
12. Tilman, D., Lehman, C. L. & Bristow, C. E. Diversity-stability relationships: statistical inevitability or ecological consequence? *Am. Nat.* **151**, 277–282 (1998).
13. Tilman, D. Biodiversity: population versus ecosystem stability. *Ecology* **77**, 350–363 (1996).
14. Tilman, D. The ecological consequences of changes in biodiversity: a search for general principles. *Ecology* **80**, 1455–1474 (1999).
15. Hastings, A. Disturbance, coexistence, history, and competition for space. *Theor. Popul. Biol.* **18**, 363–373 (1980).
16. Armstrong, R. A. & McGehee, R. Competitive exclusion. *Am. Nat.* **115**, 151–170 (1980).
17. Huisman, J. & Weissing, F. J. Biodiversity of phytoplankton by species oscillations and chaos. *Nature* **402**, 407–410 (1999).
18. Tilman, D. *Resource Competition and Community Structure* (Monographs in Population Biology, Princeton Univ. Press, 1982).
19. Loreau, M. Biodiversity and ecosystem functioning: a mechanistic model. *Proc. Natl Acad. Sci. USA* **95**, 5632–5636 (1998).
20. Tilman, D., Lehman, C. L. & Thomson, K. T. Plant diversity and ecosystem productivity: theoretical considerations. *Proc. Natl Acad. Sci. USA* **94**, 1857–1861 (1997).
21. Huston, M. A. Hidden treatments in ecological experiments: re-evaluating the ecosystem function of biodiversity. *Oecologia* **110**, 449–460 (1997).
22. Aarssen, L. W. High productivity in grassland ecosystems: effected by species diversity or productive species? *Oikos* **80**, 183–184 (1997).
23. Daily, G. C. *Nature's Services: Societal Dependence on Natural Ecosystems* (Island, Washington DC, 1997).
24. Fehr, W. H. *Genetic Contributions to Yield Gains of Five Major Crop Plants* (Crop Science Society of America, Madison, WI, 1984).
25. Roelfs, A. P. Genetic control of phenotypes in wheat stem rust. *Annu. Rev. Phytopathol.* **26**, 351–367 (1988).
26. Maddison, A. *Monitoring the World Economy 1820–1992* (Development Centre of the Organization for Economic Co-operation and Development, Paris, 1995).



Getting the measure of biodiversity

Andy Purvis* & Andy Hector†

*Department of Biology and †NERC Centre for Population Biology, Imperial College, Silwood Park, Ascot, Berkshire SL5 7PY, UK

The term 'biodiversity' is a simple contraction of 'biological diversity', and at first sight the concept is simple too: biodiversity is the sum total of all biotic variation from the level of genes to ecosystems. The challenge comes in measuring such a broad concept in ways that are useful. We show that, although biodiversity can never be fully captured by a single number, study of particular facets has led to rapid, exciting and sometimes alarming discoveries. Phylogenetic and temporal analyses are shedding light on the ecological and evolutionary processes that have shaped current biodiversity. There is no doubt that humans are now destroying this diversity at an alarming rate. A vital question now being tackled is how badly this loss affects ecosystem functioning. Although current research efforts are impressive, they are tiny in comparison to the amount of unknown diversity and the urgency and importance of the task.

To proceed very far with the study of biodiversity, we need to pin the concept down. We cannot even begin to look at how biodiversity is distributed, or how fast it is disappearing, unless we can put units on it. However, any attempt to measure biodiversity quickly runs into the problem that it is a fundamentally multidimensional concept: it cannot be reduced sensibly to a single number^{1,2}. A simple illustration can show this. Figure 1 shows samples from the insect fauna in each of two habitats. Which sample is more diverse? At first sight it must be sample A, because it contains three species to sample B's two. But sample B is more diverse in that there is less chance in sample B that two randomly chosen individuals will be of the same species. Neither of these measures of diversity is 'wrong' — species richness and evenness are two (among many) of biodiversity's facets,

and no single number can incorporate them both without loss of information. This should not be disappointing; indeed we should probably be relieved that the variety of life cannot be expressed along a single dimension. Rather, different facets of biodiversity can each be quantified (Box 1).

Knowing the diversity (however measured) of one place, group or time is in itself more-or-less useless. But, as we shall discuss later, comparable measurements of diversity from multiple places, groups or times can help us to answer crucial questions about how the diversity arose and how we might best act to maintain it. We shall see also how the usefulness of the answers depends critically on the selection of an appropriate diversity measure. No single measure will always be appropriate (indeed, for some conservation questions, no single measure can probably ever be appropriate). The choice of a good measure is complicated by the frequent

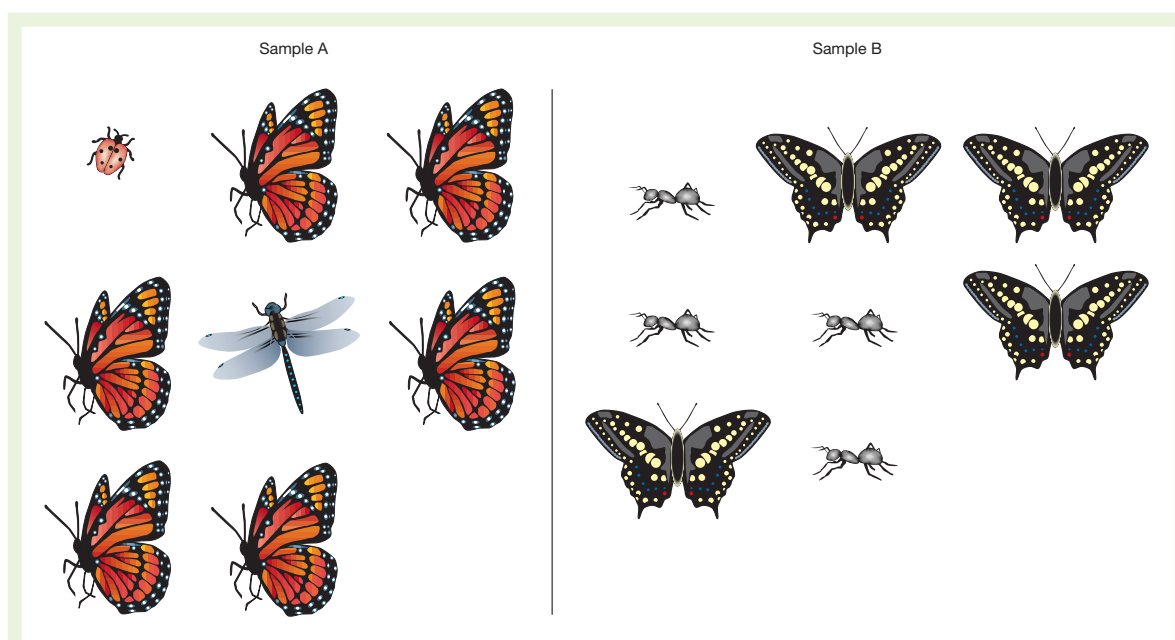


Figure 1 Two samples of insects from different locations, illustrating two of the many different measures of diversity: species richness and species evenness. Sample A could be described as being the more diverse as it contains three species to sample B's two. But there is less chance in sample B than in sample A that two randomly chosen individuals will be of the same species.

Box 1

Parts of the whole: numbers, evenness and difference

Biodiversity has a multitude of facets that can be quantified. Here we classify some commonly used measures into three conceptually different (although not orthogonal) approaches.

Numbers

The most commonly considered facet of biodiversity is species richness — the number of species in a site, habitat or clade. Species are an obvious choice of unit when trying to measure diversity. Most people have an idea what ‘species’ means and, although their ideas differ considerably (reviewed in ref. 96), there is even less commonality about other levels in the taxonomic hierarchy³⁰ (Fig. 3). Many other measures are less intuitive, and have arisen only through appreciation of limitations of measures of species richness. Species are also sensible units to choose from a biological perspective: they keep their genes more or less to themselves, and to that extent have independent evolutionary trajectories and unique histories. The current ‘best guess’¹⁷ is that there are around 14 million species, but this is very much a provisional working figure. Regions with many species, especially endemic species, are sometimes called hotspots⁹⁷.

Species and regions differ in their number of populations. Populations of a given species, if defined on the basis of limited gene flow among them, will evolve to an extent independently. Each population contributes additional diversity. The number of genetic populations in the world has been estimated to lie between 1.1 and 6.6 billion⁶⁶.

Species or populations differ in the numbers of alleles they have at given loci. For instance, Mauritius kestrels (*Falco punctatus*) have lost over half of the alleles present historically at 12 sampled microsatellite loci⁹⁸.

Moving above the species level, higher-taxon richness is often used in studies of biodiversity, usually as a less data-demanding surrogate for species richness⁹⁹.

Evenness

A site containing a thousand species might not seem particularly diverse if 99.9% of individuals that you find belong in the same species. Many diversity indices have been developed to convey the extent to which individuals are distributed evenly among species². Most but not all combine evenness with species richness, losing information by reducing two dimensions to one. There are genetic analogues of these indices¹⁰⁰, such as heterozygosity, that incorporate both allele number and relative frequencies.

Difference

Some pairs of species (or alleles or populations) are very alike, whereas others are very different. Disparity¹⁰¹ and character diversity⁹³ are measures of phenotypic difference among the species in a sample, and can be made independent of species number. Some phenotypic characteristics might be considered more important than others, for instance the ecological diversity among species may be crucial for ecosystem functioning. Genetic variability among populations can also be measured in various ways¹⁰⁰. If populations within species differ enough either genetically or phenotypically, they may be considered to be subspecies, management units or evolutionarily significant units¹⁰²; numbers of these therefore provide estimates of difference. All these kinds of difference are likely to be at least partly reflected by the phylogenetic diversity¹⁰³ among organisms, which is estimated as the sum total of the branch lengths in the phylogeny (evolutionary tree) linking them.

Sample in different places, and you will find different things. This spatial turnover itself has many facets² (for example, beta diversity, gamma diversity and numbers of habitat types), and important consequences for any attempt to conserve overall diversity (see review by Margules and Pressey, pages 243–253, and refs 104, 105). Likewise, temporal turnover¹⁰⁶ is the extent to which what is found changes over time.

need to use surrogates for the aspect in which we are most interested^{3,4}. Surrogacy is a pragmatic response to the frightening ignorance about what is out there. Some recent discoveries highlight just how much we probably still do not know.

The growing biosphere

Technological advances and the sense of urgency imparted by the rate of habitat loss are combining to yield discoveries at an incredible rate. This may seem surprising, given that expedition accounts of natural historians from the 18th and 19th centuries conjure up images of discovery on a grand scale that seemingly cannot be matched today — look in the rocks ... a new fossil mammal; look in the lake ... a new fish genus; look on the dinner plate ... a new species of bird. Finding new large vertebrates nowadays is indeed newsworthy, but a new species of large mammal is still discovered roughly every three years⁵ and a new large vertebrate from the open ocean every five years⁶. And most organisms are much smaller than these are. An average day sees the formal description of around 300 new species across the whole range of life, and there is no slowdown in sight. Based on rates of discovery and geographical scaling-up, it seems that the roughly 1.75 million described species of organism may be only around 10% of the total⁷.

It is not only new species that are discovered. Cycliophora and Loricifera are animal phyla (the level just below kingdom in the taxonomic hierarchy) that are new to science in the past 20 years⁸. Within the Archaea, the discovery of new phylum-level groups proceeds at the rate of more than one a month⁹. The physical limits of the biosphere have been pushed back by the recent discovery of microbial communities in sedimentary and even igneous rocks over 2 km

below the surface; these subsurface lithoautotrophic microbial ecosystems (termed SLiMEs) may have persisted for millions of years without any carbon from the surface¹⁰. Controversy surrounds another proposed discovery: whether or not the 100-nm-diameter nanobacteria found in, among other places, kidney stones are living organisms¹¹. At an even smaller scale, genomes provide fossils that indicate great past retroviral diversity¹². Genomes have also been found to provide habitats for many kinds of genetic entity — transposable elements — that can move around and replicate themselves. Such elements can provide important genetic variation to their hosts, can make up more than half of the host’s genome¹³, and have life histories of their own¹⁴.

There are two other ways in which the biosphere can perhaps be said to be growing. The first is that the rate at which taxonomists split one previously recognized species into two or more exceeds the rate at which they lump different species together, especially in taxa that are of particular concern to conservationists (for example, platyrrhine primates¹⁵). Part of the reason is the growing popularity of one way of delimiting species — the phylogenetic species concept (PSC)¹⁶ — under which taxa are separate species if they can be diagnosed as distinct, whether on the basis of phenotype or genotype. If the PSC becomes widely applied — which is a controversial issue¹⁷ — then the numbers of ‘species’ in many groups are sure to increase greatly¹⁸ (although the amount of disparity will barely increase at all).

A second way in which the catalogue of diversity is growing is that computer databases and the Internet are making the process of information gathering more truly cumulative than perhaps ever before. Some existing sites serve to provide examples of the information already available: not just species lists (<http://www.sp2000.org/>), but

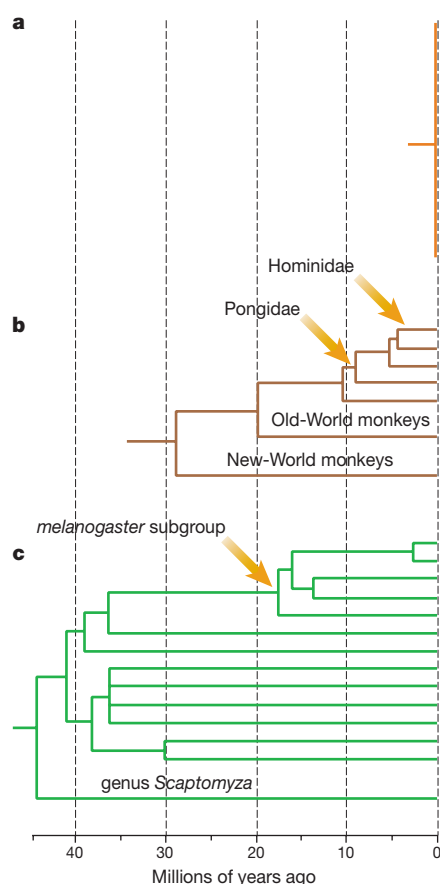


Figure 2 Taxonomic boundaries are not comparable among major groups. **a**, Fourteen species in nine genera representative of cichlid fish in Lake Victoria. **b**, Seven species representative of several families in anthropoid primates. **c**, Thirteen species representative of a single genus, *Drosophila*. Figure reproduced from ref 30, with permission.

also maps of the geographical ranges of species (<http://www.gisbau.uniroma1.it/amd/homepage.html>), information on conservation status of species (<http://www.wcmc.org.uk>), bibliographies (http://etweb.lscf.ucsb.edu/bfv/bfv_form.html), data on molecular sequence (<http://www.ebi.ac.uk/> and <http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html>), data on phylogenetic position (<http://phylogeny.arizona.edu/tree/phylogeny.html> and <http://herbaria.harvard.edu/treebase/>), information on the stratigraphic range of species (<http://ibs.uel.ac.uk/ibs/palaeo/benton/> and <http://www.nceas.ucsb.edu/~alroy/nafmtd.html>) and much more. Although the terabytes of information already stored constitute only a small drop in the ocean, the next two sections show how much can be seen in that droplet about the distribution of biodiversity among evolutionary lineages and through time.

Learning from the tree of life

The ongoing explosion of phylogenetic studies not only provides an ever-clearer snapshot of biodiversity today, but also allows us to make inferences about how the diversity has come about^{19–21}. (For an ecological perspective, see review by Gaston, pages 220–227.) Phylogenies give key information that is not available from species lists or taxonomies. They detail the pattern of nested relationships among species, and increasingly provide at least a rough timescale even without reliance on a molecular clock²². These new phylogenies are pushing back the origins of many groups to long before their earliest

known fossils. The palaeontological record indicates a Cambrian explosion of phyla around 540 million years (Myr) ago, but sequences suggest a more gradual series of splits around twice as old²³. Likewise, many orders of mammals and birds are now thought to have originated long before the end-Cretaceous extinction^{24,25}, which occurred 65 Myr ago and which was thought previously to have been the signal for their radiation. If the new timescale can be trusted²⁶, these findings present a puzzle and a warning. The puzzle is the absence of fossils. Why have we not found traces of these lineages in their first tens or even hundreds of millions of years? It seems likely that the animals were too small or too rare, with the sudden appearance in the rocks corresponding to an increase in size and rise to ecological dominance²⁷. The warning is that current biodiversity is in a sense greater than we had realized. Major lineages alive today represent more unique evolutionary history than previously suspected — history that would be lost with their extinction.

Analysis of the shape of phylogenies has shown that lineages have differed in their potential for diversification. Darwin²⁸ had noted that species in species-rich genera had more subspecific varieties, and subtaxa within taxa are often distributed very unevenly²⁹, as Fig. 2 illustrates for eutherian species. But these taxonomic patterns can be taken at face value only if taxa are comparable, which they may not be. For example, species-rich groups may simply be older, and it is clear that workers on different groups currently place taxonomic boundaries in very different places³⁰ (Fig. 3). Phylogenies allow comparison of sister clades — each other's closest relatives — which by definition are the same age. Time and again, species are distributed too unevenly for simple null models to be tested in which all species have the same chances of diversifying^{31,32}.

What are the species-rich groups 'doing right'? Many explanations fall broadly into two types. Key innovation hypotheses³³ posit the evolution of some trait that permits its bearers to gain access to more resources or be more competitive than non-bearers. Examples include phytophagy in insects³⁴ and high reproductive rate in mammals³⁵. Other hypotheses focus on traits that facilitate the evolution of reproductive isolation — speciation — without necessarily increasing the fitness of bearers. Sexual selection³⁶ and range fragmentation³⁷ are examples of this kind. These two types can be contrasted as 'bigger cake' and 'thinner slices' explanations, although some traits may act in both ways (for example, body size^{38,39}); another way to split them is to view diversity as 'demand-driven' (niches are waiting to be filled, and differentiation leads to speciation) or 'supply-driven' (speciation occurs unbidden, with differentiation arising through character displacement). Statistical testing of many key innovation hypotheses is hampered by a lack of replication — often, the trait in question is unique, and all that can be done is to model the trait's evolution to assess how well it fits the scenario⁴⁰. When characters have evolved multiple times in independent lineages, sister clades provide automatic matched pairs for hypothesis testing (although other phylogenetic approaches are also available^{41,42}). Comparing sister clades (the procedure used in most of the examples above) avoids two problems that otherwise cloud the issue. First, taxa may not be comparable (Fig. 3), and second, they are not statistically independent — related clades inherit their traits from common ancestors, so are pseudoreplicates⁴³. Nonetheless, there is ongoing debate about the role and limitations of phylogenetic tests for correlates of species richness^{44,45}.

Temporal patterns in biodiversity

Is biodiversity typically at some equilibrium level, with competition setting an upper limit, or do mass extinctions occur so regularly that equilibrium is never reached? And, with one eye on the future prospects for biodiversity, how quickly does diversity recover from mass extinctions? Palaeontologists have addressed these questions at many scales, from local to global. For the global view, the data come from huge compendia of stratigraphic ranges of

taxonomic families (see, for example, refs 46, 47), led by Sepkoski's ground-breaking efforts, and made possible by the development of computer databases. There are more families now than ever before, and a model of exponential growth provides a good overall fit to the numbers of families through time, suggesting expansion without limit and no major role for competition in limiting diversity⁴⁸. But a significantly better fit is provided by a set of three logistic curves, each with a different carrying capacity, punctuated by mass extinction events⁴⁹. Leaving aside the thorny issue of multiplicity of tests and the big question of why the three carrying capacities are different, there may be a perceptual problem at play here. Families do not arise overnight: they are the result of speciation and a lot of time. Consequently, exponential growth at the species level might appear like logistic growth at higher levels⁵⁰. This problem of perception is a recurrent one in palaeontology. For instance, good evidence that biodiversity is often near equilibrium comes from the fact that extinction events are commonly followed by higher than normal rates of diversification⁴. However, the peak of origination rates of genera and families is not straight after the extinction peak. Instead, there is a 10-Myr time-lag throughout the fossil record, implying a lag phase before diversification occurs⁵¹. But could the same pattern arise if speciation rates rose immediately in response to the extinction, but the new lineages are given generic or familial rank only after being around for some time? This scenario would predict (incorrectly) that family diversification rates would take longer to respond than generic rates, so cannot be the whole story, but it

highlights the difficulties of taking taxonomic patterns at face value. Neontologists may face much the same problem with species: taxonomists tend to recognize bird lineages as species if they are older than 2.8 Myr but not if they are younger than 1.1 Myr (ref. 52), so apparent logistic growth in species numbers through time within bird genera⁵³ might be expected even without a slow-down of cladogenesis.

The patchy nature of the known fossil record means that some taxa in some places at some times can be studied in much greater detail than is possible for the biota as a whole. Studies at these smaller scales can analyse the record at the species level, within a region or biome, and can better control for problems such as incomplete and uneven sampling^{54,55}. Such studies find a range of answers: communities may show an equilibrium diversity^{55,56}, an increasing geographical turnover⁵⁷, or radiation punctuated by mass extinction⁵⁸. This may be a more appropriate spatial scale at which to look for equilibrium, as the units have a greater chance of interacting⁵⁹.

The temporal pattern of disparity is also of great interest. Does difference accumulate gradually and evenly as lineages evolve their separate ways, or is evolutionary change more rapid early in a group's history, as it stakes its claim to a new niche? Information from living and fossil species and phylogenies can be combined with statistical models^{41,60,61} to answer this question, although so far relatively little work has combined palaeontological and neontological data. Rates of morphological and taxic diversification are often incongruent, or even uncoupled⁶¹, again highlighting that there is

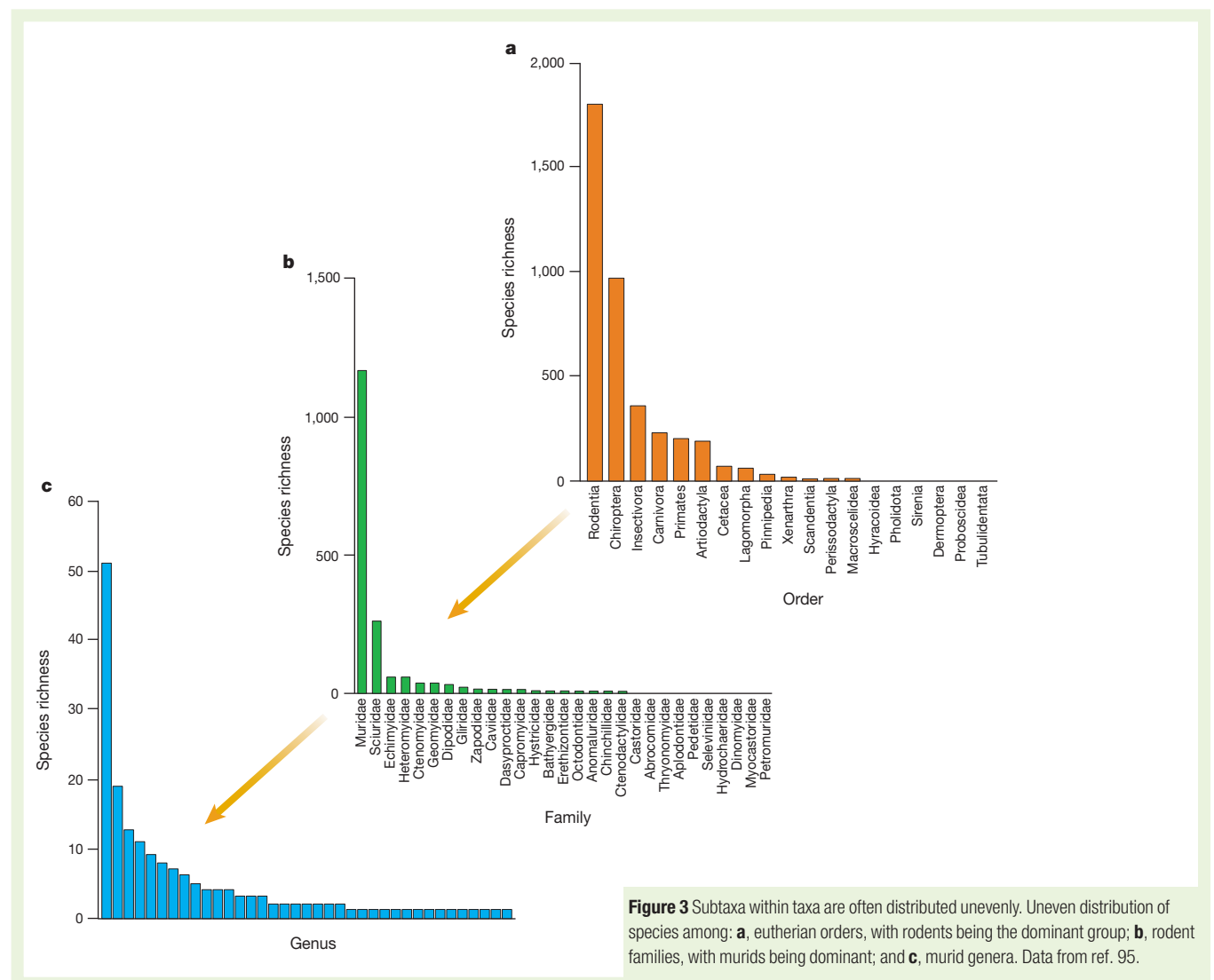


Figure 3 Subtaxa within taxa are often distributed unevenly. Uneven distribution of species among: **a**, eutherian orders, with rodents being the dominant group; **b**, rodent families, with murids being dominant; and **c**, murid genera. Data from ref. 95.

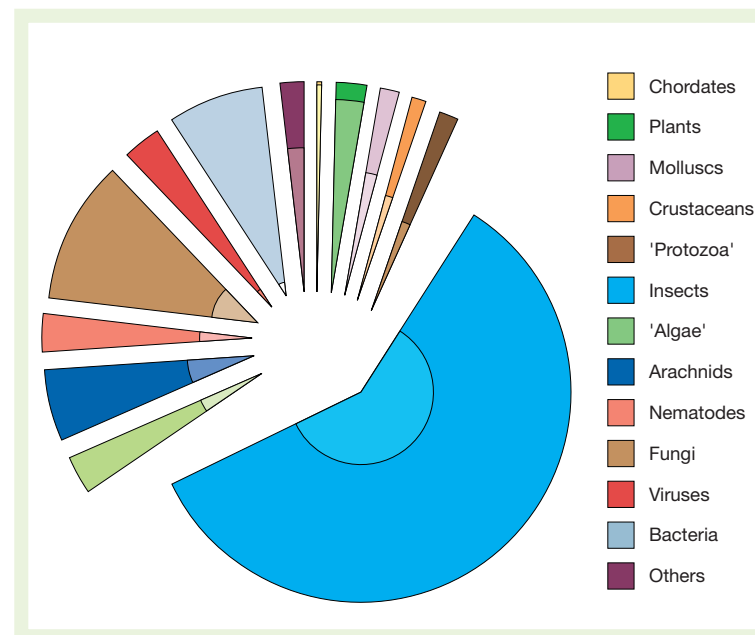


Figure 4 Species richness in major groups of organisms. The main 'pie' shows the species estimated to exist in each group; the hatched area within each slice shows the proportion that have been formally described. Data from ref. 7.

more to biodiversity than numbers of taxa. At present, it is hard to tell under what circumstances disparity precedes, or perhaps drives, species richness, and when the reverse applies. Different models can give very similar patterns of diversity and disparity over time⁶⁰, and detailed studies at smaller scale^{62,63} may provide the greatest chance of an answer.

The shrinking biosphere

What about human impacts on biodiversity? A simple calculation shows that recent rates of species losses are unsustainable. If there are 14 million species at present⁷, then each year the tree of life grows by an extra 14 Myr of branch length. The average age of extant species is nearly 5 Myr (in primates and carnivores anyway, and species in most other groups probably tend to be older rather than younger). So the tree can 'afford' at most about three species extinctions per year without shrinking overall. There have been roughly this many documented species extinctions per year since 1600⁶⁴, and most extinctions must have passed us by. The rate has been increasing too: the last century saw the end of 20 mammalian species alone, a pruning of the mammalian tree that would take at least 200 centuries to redress.

Estimates of current and future rates of loss make even more sobering reading. The rate at which tropical forest — probably the habitat for most species — is lost is about 0.8% to 2% per year⁶⁵ (call it 1% for the purpose of this example). We must expect about 1% of the tropical forest populations to be lost with it, a figure that may be as high as 16 million populations per year, or one every two seconds⁶⁶. Most species have multiple populations, so rates of species loss will obviously be much lower. They are most commonly estimated through species–area relationships⁶⁵, although other approaches are used too⁶⁷. Wilson⁶⁸ famously used the species–area relationship to estimate an annual extinction rate of 27,000 species — one species every twenty minutes. This and similar estimates have attracted criticism but recent work^{67,69,70} has shown that levels of species endangerment are rising in line with species–area predictions, provided the analysis is conducted at the appropriate scale. What are the implications of such rapid pruning for the tree of life? Simulations in which species are wiped out at random⁷¹ indicate that most of the phylogenetic diversity would survive even a major extinction: up to 80% of the branch length could survive even if 95% of the species were lost. This result assumes extinction to befall species at random; scenarios of non-random extinction can have very different outcomes⁷². The current crisis, like previous mass

extinctions, is highly non-random^{73–76}, with related twigs on the tree tending to share the same fate. This selectivity greatly reduces the ability of the phylogenetic hierarchy to retain structure in the face of a given severity of species extinction^{77,78}.

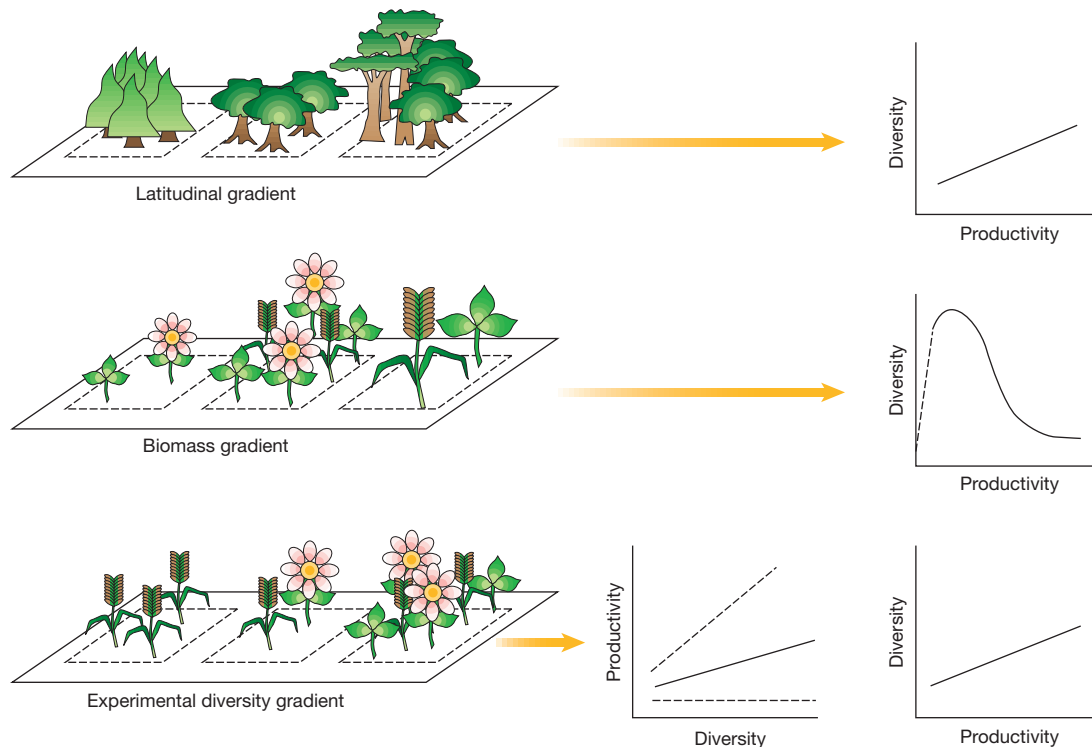
But how much structure is needed? Imagine if the only function of this article was the transfer of information. Many of the words could be deleted and you would still get the message. It would (we hope) be less pleasant to read. Similarly, for many people we need biodiversity because we like it; it should be conserved just as we conserve Mozart concertos and Van Gogh paintings⁷⁹. But how many words could you delete before the meaning starts to get lost? Recently, ecologists have begun asking similar questions about our environment.

Biodiversity and the stability and functioning of ecosystems

How many species can we lose before we start to affect the way ecosystems function? Principal environmental factors such as climate, soil type and disturbance^{80,81} strongly influence ecosystem functioning, but likewise organisms can affect their environment⁸². Some of the first ideas on how biodiversity could affect the way ecosystems function are attributable to Darwin and Wallace^{28,83}, who stated that a diverse mixture of plants should be more productive than a monoculture. They also suggested the underlying biological mechanism: because coexisting species differ ecologically, loss of a species could result in vacant niche-space and potential impacts on ecosystem processes. Defining ecological niches is not straightforward, but Darwin and Wallace's hypothesis, if correct, provides a general biological principle which predicts that intact, diverse communities are generally more stable and function better than versions that have lost species. Recent experimental evidence (reviewed by Chapin *et al.*, pages 234–242, and McCann, pages 228–233), although pointing out important exceptions, generally supports this idea. Compared with systems that have lost species, diverse plant communities often have a greater variety of positive and complementary interactions and so outperform any single species^{84,85}, and have more chance of having the right species in the right place at the right time. This last 'sampling effect' mechanism has prompted much debate on the design, analysis and interpretation of experiments that aim to manipulate biodiversity⁸⁶. Although the sampling effect is biological in part — it requires both differences between species and an ecological mechanism making some species more abundant than others — the probabilistic component (more diverse communities have a greater chance of containing a species with particular properties) has made it controversial. Nevertheless, loss of species with key

Box 2

Plant diversity and productivity at different scales



For plants, the relationship between diversity and productivity changes with scale^{107,108}. At global scales (panel **a** in the figure above), from high latitudes to the tropics, plant diversity in large areas may be positively related to increasing productivity. At regional scales (**b**), plant diversity in small plots is frequently negatively related to increasing productivity, often as part of a larger unimodal 'hump-shaped' distribution of diversities. Numbers of species correlate with several factors including the size and hence number of individual plants sampled, spatial heterogeneity, and competitive exclusion as

productivity increases. Experimental manipulations of plant diversity within habitats (**c**) reveal that, although relationships vary, productivity tends to increase with diversity owing to increasing complementary or positive interactions between species and the greater likelihood of diverse communities containing a highly productive species. In manipulation experiments, biodiversity is the explanatory variable and productivity the response, whereas in observational studies the relationship is usually viewed the other way round as illustrated here for all three cases.

traits, as in the sampling effect, is not restricted to ecological experiments: logging, fishing, trapping and other harvesting of natural resources frequently remove particular organisms, often including dominant species.

Although 95% of experimental studies support a positive relationship between diversity and ecosystem functioning, many have found that only 20–50% of species are needed to maintain most biogeochemical ecosystem processes⁸⁷. Do the other, apparently redundant, species have a role to play over longer timescales, providing insurance against environmental change? We need to know. Biodiversity can also impact ecological processes such as the incidence of herbivory and disease, and the resistance of communities to invasion. Once again, although exceptions exist, in experiments which manipulate diversity directly, communities with more species are often more resistant to invasion^{88,89}, probably for the same reason that they are more productive. Diversity of one group of organisms can also promote diversity of associated groups, for example between mycorrhizas and plants⁹⁰ or plants and insects⁸⁸.

The study of the relationship between biodiversity and ecosystem processes has made rapid progress in the past decade, and is proving an effective catalyst for linking the ecology of individuals, communities and ecosystems. Some general, although not universal, patterns are emerging as theory and experiment progress together⁹¹. We have a good understanding of the underlying causes, where we see both

agreement and differences in experimental results. Nevertheless, this work represents only a first general approach to the subject; many issues remain outstanding and other areas are as yet uninvestigated. First, do these short-term and small-scale experiments in field plots reveal the full effects of diversity, and how do we scale up in time and space⁹²? Second, although we know that local extinction is often not random, many recent experiments compare the performance of communities differing in the presence or absence of a random set of species. How adequate is this model? Third, how will species loss interact with other components of global change such as rising CO₂? Darwin and Wallace observed that niche differentiation could cause changing diversity to have consequences for ecosystem processes, but the magnitude of these effects could depend crucially on the exact mechanism of coexistence. Finally, how do we integrate these new within-habitat relationships between diversity and ecosystem processes with large-scale patterns in biodiversity and environmental parameters, as reviewed by Gaston on pages 220–227 of this issue? Box 2 suggests one way in which the relationship between plant diversity and productivity could vary with scale.

Challenges and prospects

Recent years have seen exciting advances in our knowledge of biodiversity, our identification of factors that have shaped its evolution and distribution, and our understanding of its importance. But we

can see only a small, probably atypical, part of the picture (Fig. 4). A detailed view is emerging of birds, mammals, angiosperms, and shallow-sea, hard-bodied invertebrates, but much less is known about most of the rest of life. How far are we justified in generalizing from the groups we know well to biodiversity as a whole? This is a crucial question, for instance in the choice of protected areas (see review by Margules and Pressey, pp. 243–253). There is no short cut — we need more basic information about more groups; and not just species lists, but who does what and with whom.

A related point is that biodiversity cannot be reduced to a single number, such as species richness. This is a real problem for biologists, because a single number is often what policy-makers want. Perhaps it will be possible to go part way if the many indices (Box 1) are intercorrelated, as some certainly are^{93,94}. The stronger the correlations, the more reasonable it will be to reduce multiple measures to a few principal components, to create dimensions of diversity. We must of course recognize — and explain to policy-makers — that combining these dimensions into a single number would be arbitrary. We must not make the mistake of thinking or claiming that maintaining, say, species richness of a particular taxon is the same as conserving overall biodiversity. To revisit an earlier metaphor, conserving one population of every species is rather like having one of each note in the Mozart concerto.

Two themes running through this review pertain to scale. The first is that the study of biodiversity is becoming an ever-bigger research enterprise. The database is (more than ever) cumulative, the analyses more ambitious and involving more people. We see this trend continuing. The second issue is whether we can study all processes at all scales. Perhaps large-scale patterns are a blunt instrument for studying the underlying processes, which may operate on much smaller scales. That said, we nonetheless would often like to scale our answers up: if a small experimental plot ‘needs’ *n* angiosperm species, or functional groups, for good ecosystem functioning, how many does 200 km² — or the planet — ‘need’?⁹². Given the speed at which we are pruning the tree of life, we need good answers quickly. □

- Whittaker, R. H. Evolution and measurement of species diversity. *Taxon* **21**, 213–251 (1972).
- Magurran, A. E. *Ecological Diversity and its Measurement* (Croom Helm, London, 1988).
- Balmford, A., Green, M. J. B. & Murray, M. G. Using higher-taxon richness as a surrogate for species-richness: I. Regional tests. *Proc. R. Soc. Lond. B* **263**, 1267–1274 (1996).
- Sepkoski, J. J. Jr Rates of speciation in the fossil record. *Proc. R. Soc. Lond. B* **353**, 315–326 (1998).
- Pine, R. H. New mammals not so seldom. *Nature* **368** (1994).
- Paxton, C. G. M. A cumulative species description curve for large open water marine animals. *J. Mar. Biol. Assoc.* **78**, 1389–1391 (1998).
- Hawthornthwaite, D. L. & Kalin-Arroyo, M. T. in *Global Biodiversity Assessment* (ed. Heywood, V. H.) 107–191 (Cambridge Univ. Press, Cambridge, 1995).
- Funch, P. & Kristensen, R. M. Cyclophora is a new phylum with affinities to Entoprocta and Ectoprocta. *Nature* **378**, 711–714 (1995).
- Fuhrman, J. A. & Campbell, L. Marine ecology: microbial microdiversity. *Nature* **393**, 410–411 (1998).
- Gross, M. *Life on the Edge* (Plenum, New York, 1998).
- Abbott, A. Battle lines drawn between ‘nanobacteria’ researchers. *Nature* **401**, 105 (1999).
- Tristem, M. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *J. Virol.* **74**, 3715–3730 (2000).
- Kidwell, M. G. & Lisch, D. Transposable elements as sources of variation in animals and plants. *Proc. Natl Acad. Sci. USA* **94**, 7704–7711 (1997).
- Goddard, M. R. & Burt, A. Recurrent invasion and extinction of a selfish gene. *Proc. Natl Acad. Sci. USA* **96**, 13880–13885 (1999).
- Rylands, A. B., Mittermeier, R. A. & Luna, E. R. A species list for the New World Primates (Platyrrhini): distribution by country, endemism, and conservation status according to the Macle-Lande system. *Neotrop. Primates* **35**, 113–160 (1995).
- Cracraft, J. Species concepts and speciation analysis. *Curr. Ornithol.* **1**, 159–187 (1983).
- Avise, J. C. & Wollenberg, K. Phylogenetics and the origin of species. *Proc. Natl Acad. Sci. USA* **94**, 7748–7755 (1997).
- Hanken, J. Why are there so many new amphibian species when amphibians are declining? *Trends Ecol. Evol.* **14**, 7–8 (1999).
- Harvey, P. H., Leigh Brown, A. J., Maynard Smith, J. & Nee, S. *New Uses for New Phylogenies* (Oxford Univ. Press, Oxford, 1996).
- Nee, S., Barraclough, T. G. & Harvey, P. H. in *Biodiversity: A Biology of Numbers and Difference* (ed. Gaston, K. J.) 230–252 (Blackwell Science, Oxford, 1996).
- Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
- Sanderson, M. J. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* **14**, 1218–1231 (1997).
- Wang, D. Y.-C., Kumar, S. & Hedges, S. B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B* **266**, 163–171 (1999).
- Bromham, L., Phillips, M. J. & Penny, D. Growing up with dinosaurs: molecular dates and the mammalian radiation. *Trends Ecol. Evol.* **14**, 113–118 (1999).
- Cooper, A. & Penny, D. Mass survival of birds across the Cretaceous-Tertiary boundary: molecular evidence. *Science* **275**, 1109–1113 (1997).
- Foot, M., Hunter, J. P., Janis, C. M. & Sepkoski, J. J. Evolutionary and preservational constraints on the origins of biological groups: divergence times of eutherian mammals. *Science* **283**, 1310–1314 (1999).
- Cooper, A. & Fortey, R. Evolutionary explosions and the phylogenetic fuse. *Trends Ecol. Evol.* **13**, 151–156 (1998).
- Darwin, C. *On the Origin of Species by Means of Natural Selection* (Murray, London, 1859).
- Dial, K. P. & Marzluff, J. M. Nonrandom diversification within taxonomic assemblages. *Syst. Zool.* **38**, 26–37 (1989).
- Avise, J. C. & Johns, G. C. Proposal for a standardized temporal scheme of biological classification for extant species. *Proc. Natl Acad. Sci. USA* **96**, 7358–7363 (1999).
- Purvis, A. in *New Uses for New Phylogenies* (eds Harvey, P. H., Leigh Brown, A. J., Maynard Smith, J. & Nee, S.) 153–168 (Oxford Univ. Press, Oxford, 1996).
- Mooers, A. Ø. & Heard, S. B. Evolutionary process from phylogenetic tree shape. *Q. Rev. Biol.* **72**, 31–54 (1997).
- Heard, S. B. & Hauser, D. L. Key evolutionary innovations and their ecological mechanisms. *Hist. Biol.* **10**, 151–173 (1995).
- Mitter, C., Farrell, B. & Wiegmann, B. The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? *Am. Nat.* **132**, 107–128 (1988).
- Marzluff, J. M. & Dial, K. P. Life history correlates of taxonomic diversity. *Ecology* **72**, 428–439 (1991).
- Barraclough, T. G., Harvey, P. H. & Nee, S. Sexual selection and taxonomic diversity in passerine birds. *Proc. R. Soc. Lond. B* **259**, 211–215 (1995).
- Owens, I. P. F., Bennett, P. M. & Harvey, P. H. Species richness among birds: body size, life history, sexual selection or ecology? *Proc. R. Soc. Lond. B* **266**, 933–939 (1999).
- Gardezi, T. F. & da Silva, J. Diversity in relation to body size in mammals: a comparative study. *Am. Nat.* **153**, 110–123 (1999).
- Gittleman, J. L. & Purvis, A. Body size and species richness in primates and carnivores. *Proc. R. Soc. Lond. B* **265**, 113–119 (1998).
- Sanderson, M. J. & Donoghue, M. J. Shifts in diversification rate with the origin of angiosperms. *Science* **264**, 1590–1593 (1994).
- Pagel, M. Inferring evolutionary processes from phylogenies. *Zool. Scripta* **26**, 331–348 (1997).
- Kelley, S. T. & Farrell, B. D. Is specialization a dead end? The phylogeny of host use in *Dendroctonus* bark beetles (Scolytidae). *Evolution* **52**, 1731–1743 (1998).
- Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
- Rosenzweig, M. L. Colonial birds probably do speciate faster. *Evol. Ecol.* **10**, 681–683 (1996).
- Barraclough, T. G., Nee, S. & Harvey, P. H. Sister-group analysis in identifying correlates of diversification. *Evol. Ecol.* **12**, 751–754 (1998).
- Benton, M. J. *The Fossil Record 2* (Chapman & Hall, London, 1993).
- Sepkoski, J. J. A compendium of fossil marine families. *Milwaukee Publ. Mus. Contrib. Biol. Geol.* **51**, 1–125 (1982).
- Benton, M. J. Diversification and extinction in the history of life. *Science* **268**, 52–58 (1995).
- Courtillot, V. & Gaudemer, Y. Effects of mass extinctions on biodiversity. *Nature* **381**, 146–148 (1996).
- Benton, M. J. Models for the diversification of life. *Trends Ecol. Evol.* **12**, 490–495 (1997).
- Kirchner, J. W. & Weil, A. Delayed biological recovery from extinctions throughout the fossil record. *Nature* **404**, 177–180 (2000).
- Avise, J. C., Walker, D. & Johns, G. C. Speciation durations and Pleistocene effects on vertebrate phylogeography. *Proc. R. Soc. Lond. B* **265**, 1707–1712 (1998).
- Zink, R. M. & Slowinski, J. B. Evidence from molecular systematics for decreased avian diversification in the Pleistocene epoch. *Proc. Natl Acad. Sci. USA* **92**, 5832–5835 (1995).
- Marshall, C. R. in *The Adequacy of the Fossil Record* (ed. Paul, C. R. C.) 23–53 (Wiley, Chichester, 1998).
- Alroy, J. in *Biodiversity Dynamics* (eds McKinney, M. L. & Drake, J. A.) 232–287 (Columbia Univ. Press, New York, 1999).
- Alroy, J. Constant extinction, constrained diversification, and uncoordinated stasis in North American mammals. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **127**, 285–311 (1996).
- Van Valkenburgh, B. & Janis, C. in *Species Diversity in Ecological Communities* (eds Ricklefs, R. E. & Schluter, D.) 330–340 (Chicago Univ. Press, Chicago, 1993).
- Kauffman, E. G. & Fagerstrom, J. A. in *Species Diversity in Ecological Communities* (eds Ricklefs, R. E. & Schluter, D.) 315–329 (Chicago Univ. Press, Chicago, 1993).
- McKinney, M. L. in *Biodiversity Dynamics* (eds McKinney, M. L. & Drake, J. A.) 1–16 (Columbia Univ. Press, New York, 1999).
- Foot, M. in *Evolutionary Paleobiology* (eds Jablonski, D., Erwin, D. H. & Lipps, J. H.) 62–86 (Chicago Univ. Press, Chicago, 1996).
- Roy, K. & Foot, M. Morphological approaches to measuring biodiversity. *Trends Ecol. Evol.* **12**, 277–281 (1997).
- Schluter, D. Ecological causes of adaptive radiation. *Am. Nat.* **148**(Suppl.), S40–S64 (1996).
- Barraclough, T. G., Vogler, A. P. & Harvey, P. H. Revealing the factors that promote speciation. *Phil. Trans. R. Soc. Lond. B* **353**, 241–249 (1998).
- Barbault, R. & Sastrapradja, S. D. in *Global Biodiversity Assessment* (ed. Heywood, V. H.) 193–274 (Cambridge Univ. Press, Cambridge, 1995).
- May, R. M., Lawton, J. H. & Stork, N. E. in *Extinction Rates* (eds Lawton, J. H. & May, R. M.) 1–24 (Oxford Univ. Press, Oxford, 1995).
- Hughes, J. B., Daily, G. C. & Ehrlich, P. R. Population diversity: its extent and extinction. *Science* **278**, 689–692 (1997).
- Pimm, S. L. in *Conservation Science and Action* (ed. Sutherland, W. J.) 20–38 (Blackwell Science, Oxford, 1998).
- Wilson, E. O. *The Diversity of Life* (Norton, New York, 1992).
- Grelle, C. E. d. V., Fonseca, G. A. B., Fonseca, M. T. & Costa, L. P. The question of scale in threat analysis: a case study with Brazilian mammals. *Anim. Conserv.* **2**, 149–152 (1999).
- Cowlishaw, G. Predicting the pattern of decline of African primate diversity: an extinction debt from historical deforestation. *Conserv. Biol.* **13**, 1183–1193 (1999).
- Nee, S. & May, R. M. Extinction and the loss of evolutionary history. *Science* **278**, 692–694 (1997).

72. Heard, S. B. & Mooers, A. Ø. Phylogenetically patterned speciation rates and extinction risks change the loss of evolutionary history during extinctions. *Proc. R. Soc. Lond. B* **267**, 613–620 (2000).
73. Mace, G. M. & Balmford, A. in *Future Priorities for the Conservation of Mammalian Diversity* (eds Entwistle, A. & Dunstone, N.) (Cambridge Univ. Press, Cambridge, 1999).
74. Bennett, P. M. & Owens, I. P. F. Variation in extinction risk among birds: chance or evolutionary predisposition? *Proc. R. Soc. Lond. B* **264**, 401–408 (1997).
75. McKinney, M. L. Extinction vulnerability and selectivity: combining ecological and paleontological views. *Annu. Rev. Ecol. Syst.* **28**, 495–516 (1997).
76. Russell, G. J., Brooks, T. M., McKinney, M. M. & Anderson, C. G. Present and future taxonomic selectivity in bird and mammal extinctions. *Conserv. Biol.* **12**, 1365–1376 (1998).
77. McKinney, M. L. Branching models predict loss of many bird and mammal orders within centuries. *Anim. Conserv.* **1**, 159–164 (1998).
78. Purvis, A., Agapow, P.-M., Gittleman, J. L. & Mace, G. M. Nonrandom extinction risk and the loss of evolutionary history. *Science* **288**, 328–330 (2000).
79. Kunin, W. E. & Lawton, J. H. in *Biodiversity: a Biology of Numbers and Difference* (ed. Gaston, K. J.) 283–308 (Blackwell Science, Oxford, 1996).
80. MacGillivray, C. W. *et al.* Testing predictions of the resistance and resilience of vegetation subjected to extreme events. *Funct. Ecol.* **9**, 640–649 (1995).
81. Wardle, D. A., Zackrisson, O., Hörnberg, G. & Gallet, C. The influence of island area on ecosystem properties. *Science* **277**, 1296–1299 (1997).
82. Jones, C. J., Lawton, J. H. & Shachak, M. Positive and negative effects of organisms as physical ecosystem engineers. *Ecology* **78**, 1946–1957 (1997).
83. Darwin, C. & Wallace, A. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *J. Proc. Linn. Soc. Lond., Zool.* **3**, 45–62 (1858).
84. Tilman, D. Ecological consequences of biodiversity: a search for general principles. *Ecology* **80**, 1455–1474 (1999).
85. Spehn, E. M., Joshi, J., Schmid, B., Diemer, M. & Körner, C. Aboveground resource use increases with plant species richness in experimental grassland ecosystems. *Funct. Ecol.* **14** (in the press).
86. Allison, G. W. The implications of experimental design for biodiversity manipulations. *Am. Nat.* **153**, 26–45 (1999).
87. Schwartz, M. W. *et al.* Linking biodiversity to ecosystem functioning: implications for conservation ecology. *Oecologia* **122**, 297–305 (2000).
88. Knops, J. M. H. *et al.* Effects of plant species richness on invasion dynamics, disease outbreaks, insect abundance and diversity. *Ecol. Lett.* **2**, 286–294 (1999).
89. Stachowicz, J. J., Whitlatch, R. B. & Osman, R. W. Species diversity and invasion resistance in a marine ecosystem. *Science* **286**, 1577–1579 (1999).
90. Van der Heijden, M. G. A. *et al.* Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396**, 69–72 (1998).
91. Loreau, M. Biodiversity and ecosystem functioning: a mechanistic model. *Proc. Natl Acad. Sci. USA* **95**, 5632–5636 (1998).
92. Tilman, D. Diversity and production in European Grasslands. *Science* **286**, 1099–1100 (1999).
93. Williams, P. H. & Humphries, C. J. in *Biodiversity: A Biology of Numbers and Difference* (ed. Gaston, K. J.) 54–76 (Blackwell Scientific, Oxford, 1996).
94. May, R. M. in *Ecology and Evolution of Communities* (eds Cody, M. L. & Diamond, J. M.) 81–120 (Belknap, Cambridge, MA, 1975).
95. Corbet, G. B. & Hill, J. E. *A World List of Mammalian Species* (Natural History Museum, London, 1991).
96. Bisby, F. A. in *Global Biodiversity Assessment* (ed. Heywood, V. H.) 21–106 (Cambridge Univ. Press, Cambridge, 1995).
97. Myers, N., Mittermeier, R. A., Mittermeier, C. G., de Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
98. Groombridge, J. J., Jones, C. G., Bruford, M. W. & Nichols, R. A. 'Ghost' alleles of the Mauritius kestrel. *Nature* **403**, 616 (2000).
99. Williams, P. H., Humphries, C. J. & Vane-Wright, R. I. Centres of seed-plant diversity: the family way. *Proc. R. Soc. Lond. B* **256**, 67–70 (1994).
100. Mallet, J. in *Biodiversity: A Biology of Numbers and Difference* (ed. Gaston, K. J.) 13–53 (Blackwell Science, Oxford, 1996).
101. Gould, S. J. *Wonderful Life: The Burgess Shale and the Nature of History* (Norton, New York, 1989).
102. Moritz, C. Defining 'evolutionarily significant units' for conservation. *Trends Ecol. Evol.* **9**, 373–375 (1994).
103. Faith, D. P. in *Systematics and Conservation Evaluation* (eds Forey, P. L., Humphries, C. J. & Vane-Wright, R. I.) 251–268 (Clarendon, Oxford, 1994).
104. Vane-Wright, R. I., Humphries, C. J. & Williams, P. H. What to protect? Systematics and the agony of choice. *Biol. Conserv.* **55**, 235–254 (1991).
105. Pressey, R. L., Johnson, I. R. & Wilson, P. D. Shades of irreplaceability: towards a measure of the contribution of sites to a reservation goal. *Biodiv. Conserv.* **3**, 242–262 (1994).
106. Russell, G. J. in *Biodiversity Dynamics* (eds McKinney, M. L. & Drake, J. A.) 377–404 (Columbia Univ. Press, New York, 1999).
107. Grace, J. B. The factors controlling species density in herbaceous plant communities: an assessment. *Persp. Plant Ecol. Evol. Syst.* **2**, 1–28 (1999).
108. Waide, R. B. *et al.* The relationship between productivity and species richness. *Annu. Rev. Ecol. Syst.* **30**, 257–300 (1999).

Acknowledgements

We thank T. Barraclough, C. Godfray, R. Grenyer, P. Harvey, C. Humphries, N. Isaac, G. Mace, A. Minns, B. Schmid and R. Vane-Wright for discussion and comments, and NERC and the EC BIODDEPTH project for support.

Global patterns in biodiversity

Kevin J. Gaston

Biodiversity and Macroecology Group, Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK
(e-mail: k.j.gaston@sheffield.ac.uk)

To a first approximation, the distribution of biodiversity across the Earth can be described in terms of a relatively small number of broad-scale spatial patterns. Although these patterns are increasingly well documented, understanding why they exist constitutes one of the most significant intellectual challenges to ecologists and biogeographers. Theory is, however, developing rapidly, improving in its internal consistency, and more readily subjected to empirical challenge.

Biodiversity, the variety of life, is distributed heterogeneously across the Earth. Some areas teem with biological variation (for example, some moist tropical forests and coral reefs), others are virtually devoid of life (for example, some deserts and polar regions), and most fall somewhere in between. Determining why these differences occur has long been a core objective for ecologists and biogeographers. It constitutes a continuing, an important, and to many an enthralling, challenge. Indeed, the past decade has seen a veritable explosion of studies documenting broad-scale (geographical) spatial patterns in biodiversity, seeking to explain them, and exploring their implications. The reasons for this interest are twofold. First, it reflects increased opportunity provided by improvements in available data and analytical tools, the former resulting mostly from extensive collation of existing specimen and species occurrence records, the establishment of dedicated distribution-mapping schemes, and the use of remote-sensing technology (to measure vegetation and other environmental variables). Second, it reflects concern over the future of biodiversity, and the resultant need to determine its current status, to predict its likely response to global environmental change, and to identify the most effective schemes for *in situ* conservation and sustainable use. Many of these issues can be addressed satisfactorily only by resolving the historical mismatch between the fine resolution of study plots in ecological field work (typically a few square metres) and, by comparison, the poor resolution of land-use planning and models of environmental change.

A host of global patterns of spatial variation in biodiversity has been explored (Fig. 1). This includes patterns in hotspots and coldspots (highs and lows) of diversity (including comparisons between biological realms and between biogeographical regions), variation with spatial scale (for example, species–area relationships and relationships between local and regional richness) and along gradients across space or environmental conditions (for example, latitude, longitude, altitude, depth, peninsulas, bays, isolation, productivity/energy and aridity^{1,2}). Although several different levels of organization (genes to

ecosystems) of biological variation can be distinguished, most analyses of spatial variation concern biodiversity as measured by the number of species observed or estimated to occur in an area (species richness). This results from widespread recognition of the significance of the species as a biological unit, and from the practical issues of the ease and magnitude of data acquisition. Consideration of spatial variation in other measures of biodiversity, particularly those concerning the difference between entities rather than simply their numbers, has been remarkably sparse (with the possible exception of patterns in body size and morphology). Thus, although much attention has been paid to latitudinal variation in species richness, little is known about variation in the diversity of genes, individuals or populations along latitudinal gradients.

The growth of interest in broad-scale spatial variation in biodiversity has been particularly striking with regard to four areas of enquiry: latitudinal gradients in species richness, species–energy relationships, relationships between local and regional richness, and taxonomic covariance in species richness. In this review, the progress being made in each of these areas will be used to substantiate four broader cross-cutting observations about global patterns of biodiversity: respectively, that no single mechanism adequately explains all examples of a given pattern, that the patterns observed may vary with spatial scale, that processes operating at regional scales influence patterns observed at local ones, and that the relative balance of causal mechanisms means that there will invariably be variations in and exceptions to any given pattern.

Latitudinal gradients in species richness

High proportions of terrestrial and freshwater species occur in the tropics. Moving from high to low latitudes the average species richness within a sampling area of a given size increases, as has been documented for a wide spectrum of taxonomic groups (including groups as different as protists, trees, ants, woodpeckers and primates) for data across a range of spatial resolutions^{3,4}. Such gradients in species richness may be steep (for a given area, tropical assemblages are often several times more speciose than temperate ones), and have been a persistent feature of the



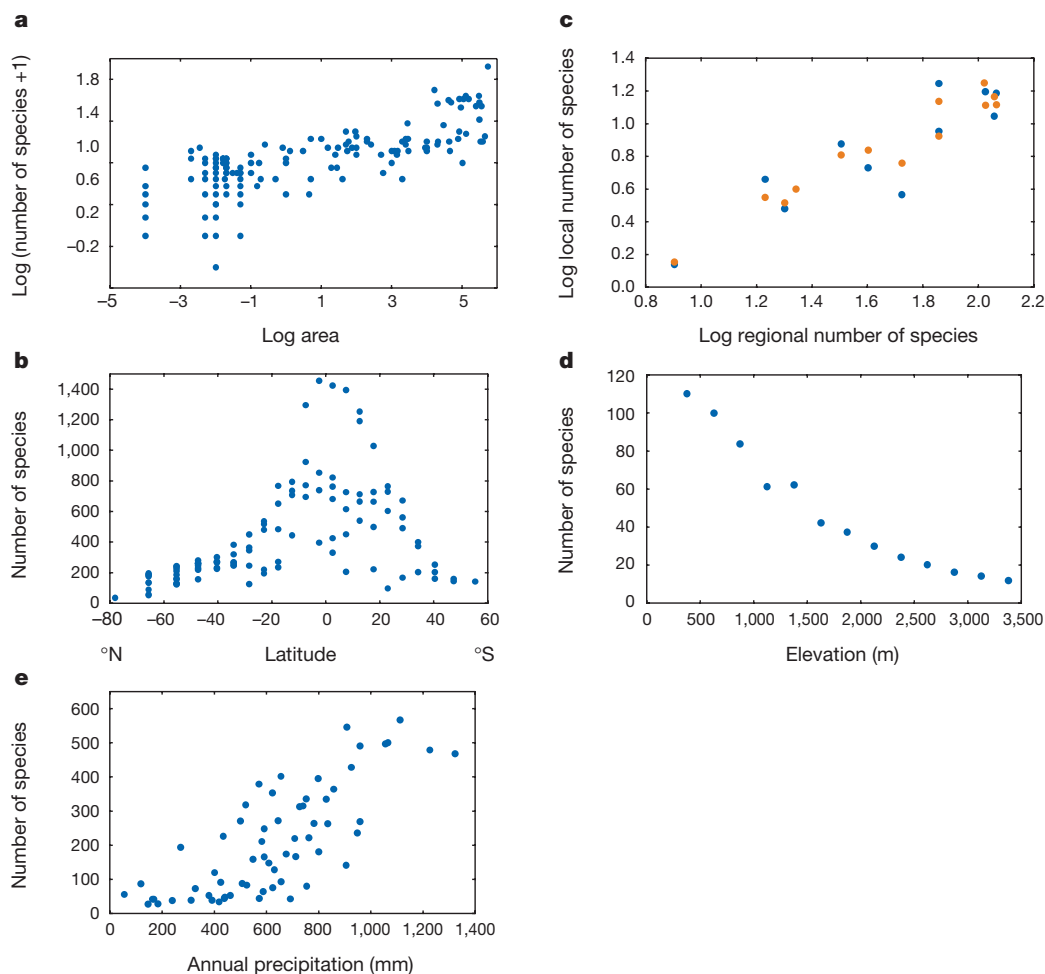


Figure 1 Spatial patterns in species richness. **a**, Species–area relationship: earthworms in areas ranging from 100 m² to >500,000 km² across Europe⁷⁶. **b**, Species–latitude relationship: birds in grid cells (~611,000 km²) across the New World⁴⁴. **c**, Relationship between local and regional richness: lacustrine fish in North America (orange circles, large lakes; blue circles, small lakes)⁶¹. **d**, Species–elevation relationship: bats in Manu National Park & Biosphere Reserve, Peru⁷⁷. **e**, Species–precipitation relationship: woody plants in grid cells (20,000 km²) in southern Africa⁷⁸.

history of biodiversity^{5,6}. In the marine environment, open-ocean pelagic and deep-sea taxa also show broad latitudinal gradients in species richness, but some debate continues to surround evidence for shallow-water systems, particularly for non-calcareous taxa⁷.

The growing number of increasingly refined analyses of latitudinal gradients in species richness has begun to suggest some important nuances to this pattern, although the extent of their generality remains uncertain. Thus, it seems that declines in richness with latitude may be faster in the Northern than in the Southern Hemisphere^{8,9}, and that peaks in richness may not lie actually at the Equator itself but some distance away^{10,11}. Although poorly documented, such latitudinal asymmetries would be unsurprising given that these exist also in contemporary climate, in historical climatic events, and in the latitudinal complexities of the geometry and area of land and ocean.

Indeed, the latitudinal gradient in species richness is a gross abstraction. Any underlying pattern is disrupted, sometimes markedly, by variation in species richness with other positional variables (for example, longitude, elevation and depth), and environmental ones (for example, topography and aridity). Thus, the detailed pattern of change with latitude depends on where one looks, reflecting the generally complex patterns of spatial variation in species richness. This indicates that consideration of latitudinal gradients in richness in isolation from other gradients might not be the most profitable way forward. In as much as latitude *per se* (and likewise other positional variables) cannot be a determinant of species richness, but only a correlate of numbers of potentially causal environmental factors, this is doubtless correct. Nonetheless, more than any other pattern the latitudinal gradient in species richness has

held an enduring fascination for biologists, particularly because of the obviously striking diversity of many tropical floras and faunas when contrasted with their counterparts at high latitudes.

The latitudinal gradient in species richness, however complex it might be, is a consequence of systematic spatial variation in the balance of speciation and the immigration of species, which add species to an area, and of the extinction and emigration of species, which take them away. For very large areas, the effects of speciation and regional or global extinction will predominate, and immigration and emigration will be less important. More than 25 different mechanisms have been suggested for generating systematic latitudinal variation in these processes², commonly emphasizing reasons as to why the tropics are highly speciose (although there is no *a priori* expectation that either tropical or temperate zones in any sense represent an ‘unusual’ condition¹²). These include explanations based on chance, historical perturbation, environmental stability, habitat heterogeneity, productivity and interspecific interactions.

Many of these mechanisms are not mutually exclusive, and others merely offer different levels of explanation. Nonetheless, to some, *en masse* they have been perceived to constitute a Gordian knot. Two recent attempts to cut it concern the importance of the physical structure of the Earth. First, null models that assume no environmental gradients, but merely a random latitudinal association between the size and placement (midpoint) of the geographical ranges of species, predict a peak of species richness at tropical latitudes¹³. This occurs because when the latitudinal extents of species in a given taxonomic group are bounded to north and south — perhaps by a physical constraint such as a continental edge or perhaps by a climatic constraint such as a critical temperature or precipitation threshold — then the

number of ways in which ranges can be distributed changes systematically between the bounds. Thus, whereas species with latitudinal midpoints midway between the bounds can extend a little or a long way before those bounds are encountered, those with midpoints close to the bounds can extend only a little way before this occurs. A null model has been wanting from discussions of latitudinal gradients in species richness. The 'mid-domain' model is thus likely to stimulate much interest. It is also likely to be most applicable for groups whose distributions are genuinely limited by a physical boundary (for example, those of large islands such as Madagascar), although its extension to two spatial dimensions is problematic, given the longitudinal variation in land and ocean area. The application of the model to other kinds of constraints is more questionable, as the position of those constraints that are recognized will be dependent on the inclusiveness of the set of species considered.

The second attempt to explain latitudinal gradients in species richness based on the physical structure of the Earth concerns the role of area (its importance has long been entertained^{14,15} and recently brought to prominence^{16,17}). The tropics have a larger climatically similar total surface area than any other ecoclimatic zone. This is because: (1) the surface area of latitudinal bands decreases towards the poles; (2) the temperature gradient between the Equator and the poles is nonlinear (the mean being relatively constant between approximately 20° N and 20° S); and (3) the regions of similar climate immediately north and south of the Equator abut. It has been contended that, for a given species richness, larger mean geographical-range sizes of species in the tropics result from the large area (which is not to be confused with any observed pattern in mean range sizes at different levels of richness), and that these translate into higher speciation rates (presuming larger ranges have higher probabilities of speciation) and lower extinction probabilities (presuming larger ranges have lower probabilities of extinction)^{16,17}. As a consequence, tropical regions have greater numbers of species than extratropical ones.

Area is almost certainly an important contributor to latitudinal gradients in species richness (indeed, area effects have a pervasive influence on patterns of biodiversity). However, tests of the 'area model' have been limited (and often tangential), and have seldom sought the signal of the influence of area on latitudinal gradients when other factors have been controlled for. Moreover, as a sole explanation the area model is insufficient. To account fully for a latitudinal gradient in species richness (rather than simply for the greater richness of the tropics) the model requires that ecoclimatic zones decline systematically in area moving from the Equator towards the poles. However, they do not do so (ecoclimatic zones at high latitudes tend to be large^{10,17}). Three possible explanations have been advanced for why a latitudinal gradient in species richness might nonetheless be expressed: (1) low productivity/energy availability at high latitudes reduces the species richness they would gain as a result of area alone^{10,17};

- (2) zonal bleeding of tropical species into extratropical regions smooths out species-richness gradients^{18,19}; and
- (3) high

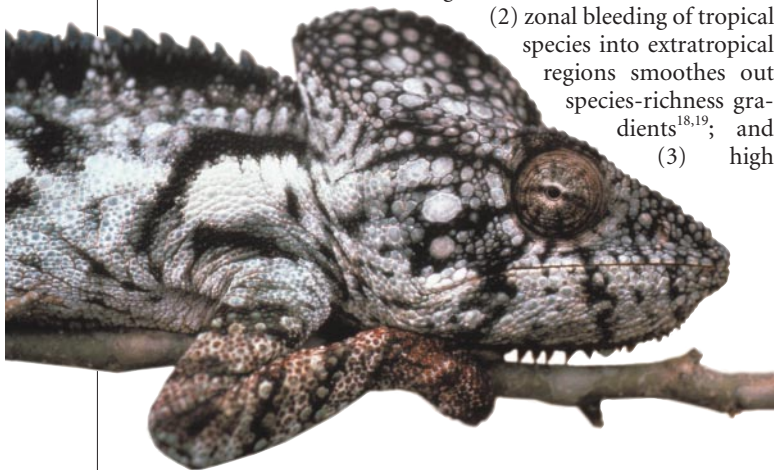
local climatic variability at high latitudes effectively increases the area of ecoclimatic zones that species can actually occupy, because it requires that individuals have broad environmental tolerances³.

The observation that area alone is insufficient as a determinant of latitudinal gradients in species richness could equally be made about almost any other factor that has been proposed as being important (although critical tests are typically lacking). This highlights an issue that has been central to much of the debate about the cause of this and other global patterns in biodiversity, namely the assumption that where a pattern is common to many taxa it must result from the same single mechanism — "wherever there is a widespread pattern, there is likely to be a general explanation which applies to the whole pattern"²⁰. To argue for a single primary cause may be to expect from ecological interactions a simplicity for which there is little evidence. There is no necessary reason why latitudinal gradients exhibited by taxa as distinct as protozoa and mammals, and in environments as structurally different as the deep sea and tropical forests, need be generated in the same way, whatever the attractions of Occam's razor. Increasingly it seems that patterns in biodiversity are likely to be generated by several contributory mechanisms^{12,21}. The strongest and most general may be those where all the different mechanisms pull in the same direction²². It is instructive that although numerous mechanisms for latitudinal gradients in species richness have been identified, and rather few processes that would oppose such a trend, no single mechanism has of itself proven sufficient.

Species–energy relationships

One factor thought to be important in modulating any effect of the physical structure of the Earth in determining latitudinal gradients in species richness is the relationship between the number of species in an area and ambient available ('usable') environmental energy. (This energy is usually estimated from models or indirectly from other variables, and often used interchangeably with 'net primary productivity'.) The form and cause of this relationship are some of the most hotly debated topics in the study of global patterns in biodiversity, with many fundamental issues as yet unresolved. Much of the discussion centres on the influence of spatial scale on observed relationships.

At a relatively local scale (spatial resolution and extent), there is a marked tendency for a general hump-shaped relationship between species richness and available energy, with species richness increasing from low to moderate levels of energy and then declining again towards high levels of energy when a sufficient range of energy values is sampled^{16,17,23}. At least across temperate to polar areas, at geographical scales there is substantial evidence for a broadly positive monotonic relationship between species richness and energy availability to be common^{10,24–33} (Fig. 2). The best correlates for plants tend to be measures of both heat and water (such as actual evapotranspiration and net primary productivity), whereas for terrestrial, and perhaps marine, animals the best correlates are measures of heat (such as mean annual temperature and potential evapotranspiration)^{28,29,34}. For example, whereas the species richness of trees in temperate Europe, eastern North America and East Asia increases with primary productivity²⁷, the richness of butterflies and birds in areas of Britain increases with the temperature during the appropriate season^{25,26}, and the species richness of amphibians, reptiles, birds and mammals in areas of North America increases with annual potential evapotranspiration (estimated as a measure of the net atmospheric energy balance, independent of water availability²⁸). The form taken by species–energy relationships at geographical scales, when extended to include subtropical and tropical areas, or at least to include the fullest range of variation in available energy (which may not be the same thing), remains unclear. There is evidence to suggest that they remain broadly positive and monotonic, that they become mildly or strongly hump-shaped, and that they begin to break down altogether^{10,32,35–37}; the answer may depend critically on the measure of energy used and the taxon concerned.



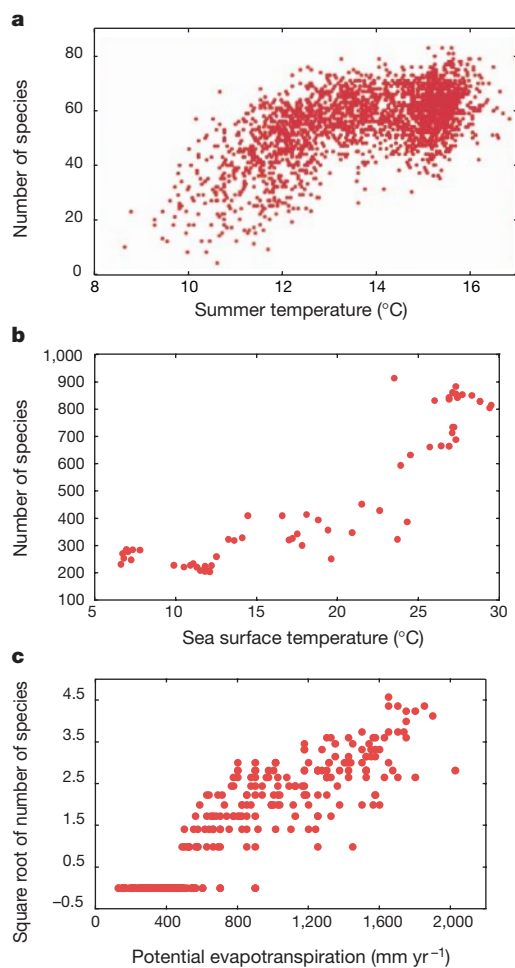


Figure 2 Species–energy relationships. **a**, Mean monthly summer temperature (°C) and richness of breeding birds in Britain (grid cells of 10 km × 10 km)³³. **b**, Mean annual sea surface temperature and richness of eastern Pacific marine gastropods (bands of 1° latitude)¹⁰. **c**, Potential evapotranspiration (mm yr⁻¹) and richness of *Epicauta* beetles (Meloidae) in North America (grid cells of 2.5° × 2.5° south of 50° N, 2.5° × 5° north of 50° N)³¹.

Any contingency of the gross or more detailed form of patterns in biodiversity on the spatial extent and dispersion of sampling units is not restricted to species–energy relationships. Indeed, the almost ubiquitous positive relationship between the numbers of species in an area and the size of that area (the species–area relationship) may itself vary in form with the absolute sizes of areas, their spatial relationships (for example, isolation), and their latitudinal position^{38,39}; this is often forgotten when attempting to control for differences in area in analyses of global patterns of biodiversity. Reconciliation of the patterns in biodiversity that are observed at different scales may provide significant insights into their determinants. If this is to be achieved, it is important to ensure that the scale of sampling and the scale of processes that are postulated to explain patterns in species richness are closely matched. One criticism of some discussion of species–energy relationships at broad scales has been that this has not been done; curiously, this has been interpreted, by different parties, as yielding species–energy relationships that may be misleadingly strong or misleadingly weak^{40–42}. Matching scales of sampling and processes is more readily achievable at local scales, and constitutes one of the most significant obstacles to testing mechanisms over broader areas.

Although other explanations have been offered, the processes

resulting in a broadly positive relationship between species richness and energy availability at geographical scales (and at low-to-moderate energy levels at more local scales) are believed to be reasonably straightforward. Greater energy availability is assumed to enable a greater biomass to be supported in an area. In turn, this enables more individual organisms to coexist, and thus more species at abundances that enable them to maintain viable populations. The result is an increase in species richness with energy availability. This assumes a basic equivalence between species in their energetic requirements at different levels of energy availability⁴³. Although there is some evidence in animal systems that average densities and body sizes of species in an area decrease as energy availability increases (that is, energy is divided more finely⁴⁴), this will tend to enhance the species–energy relationship provided these trends are sufficiently marked compared with the scaling of metabolic rate to body mass.

There are important similarities between this ‘more-individuals model’⁴⁵ and the area model as explanations of variation in species richness⁴⁴. First, both to some degree concern variation in solar energy (and water availability), with the level and availability of this energy source being important in the former case, and the spatial extent of a given level (as reflected in an ecoclimatic zone) in the latter. If ecoclimatic zones vary in available energy, then observed species–energy relationships (and those between richness and latitude) may reflect the joint effects of their area and this availability³⁷. Second, the area model assumes that area influences richness through its effect on geographical-range size, and the more-individuals hypothesis that energy influences richness through its effect on population size. There is a general, positive, interspecific relationship between total population size (or local density) and size of geographical range⁴⁶. Any factor that increases one of these variables will also be likely to increase the other. Both mechanisms therefore depend, in effect, on some factor that is posited to influence the biomass available to be worked on by the processes of speciation and extinction, which will be a product of both area and available energy per unit area^{29,47}. Presumably, it is for this reason that small areas tend to be species poor however high their energy input, whereas large areas tend to be species poor if there is low energy input.

Assuming that species–energy relations are causal and that a more-individuals model is operating, then it is unlikely that the path of causality is simple. Levels of available energy may constrain the amount of biomass that is achieved in an area, but characteristics of the biosphere, and particularly those of vegetation, are themselves known to be key influences on climate, including temperature and precipitation⁴⁸. For example, the coupling of an atmospheric model and a simple land-surface scheme has indicated that coastal deforestation in West Africa has been a significant contributor to the observed drought in the region⁴⁹; this deforestation has resulted in a number of species being threatened with extinction⁵⁰. Complex patterns of causality suggest an important connection between species–energy theory and debates over the ecosystem function of biodiversity^{51,52}.

Even accepting that paths of causality may be complex, there are some potentially significant difficulties with a more-individuals model.

1. The assumption that the number of individual organisms increases with available energy and total biomass may not apply to plants, for which there is evidence that as standing crop increases the numbers of adult individuals per unit area actually declines (and their size increases), which should tend to reduce species richness rather than increase it³⁵. However, this argument is based in large part on findings from monospecific stands of species differing substantially in their architecture, and it is unclear to what extent it generalizes to multispecies stands and systems that are structurally more similar (for example, temperate compared with tropical forests). Evidence as to how overall biomass and numbers of individuals change with species richness in animal systems is scant, even for well-known groups such as birds, and is plagued by a paucity

of strictly comparable studies from areas differing markedly in species richness.

2. Many taxa use such a small proportion of the total energy available in an area, or at least of the energy that is being measured, that it seems unlikely that detectable relationships with species richness would arise (especially given the likely magnitude of measurement errors). Thus, although species richness of birds tends to increase with available energy, avian assemblages may, directly and indirectly, commonly exploit only a small proportion of the primary production in a locality. (The avian community of the forested watersheds of the Hubbard Brook Experimental Forest has an average ingestion rate which represents 0.17% of ecosystem net annual productivity⁵³.)

3. In its simplest form, the more-individuals model ignores the likely effects of temporal variance in energy levels on species richness. High average levels of energy may not result in large numbers of species if they are accompanied by high temporal variability in those levels. The relationship between levels of available energy and their variance may be broadly different between some terrestrial and marine systems (negative in the former, positive in the latter), perhaps explaining why even at very broad spatial scales high richness may not be associated with high productivity in marine systems³⁷.

4. At regional scales, levels of species richness have not been produced directly by present environmental conditions, as processes of speciation and extinction do not operate on these timescales. If the more-individuals model is to apply this must mean that present environmental conditions are a good proxy for past ones, or at least of relative differences in the conditions in different areas.

Alternatives to the more-individuals model have been advanced to explain positive species–energy relationships. These have been based particularly on variation with energy in levels of constraints on geographical ranges, specialization, population growth rates and number of trophic levels⁴⁵. Foremost is the idea that the relationships may reflect physiological constraints on the distribution of species, with energy availability capturing factors that limit distributions as a result of metabolic considerations³⁰.

In the absence of strong support for any of these alternative explanations, difficulties with the more-individuals model fuel growing speculation that at least some species–energy relationships may not be causal, and that energy availability may often be only a covariate of some other factor that is actually driving species richness. Bird richness may, for example, be responding to a second-order effect of greater vegetational complexity with increased available environmental energy. Likewise, recent work has shown that whereas sea surface temperature explained nearly 90% of geographical variation in planktonic foraminiferal diversity throughout the Atlantic Ocean, this temperature was also correlated with temperatures at different depths. This indicates that the diversity may be controlled by the physical structure of the near-surface ocean and not directly by available energy³².

Continuing with this theme, there has been debate as to the respective roles of contemporary levels of energy and of historical factors in generating global patterns of tree species richness in moist forests. The debate has centred on the extent to which differences in richness between continents and between latitudes result from variation in annual actual evapotranspiration (a good, but not universal, predictor of primary productivity) or from long-term evolutionary and geographical processes^{40–42}. The practical constraints on conducting experiments at relevant scales mean that differentiating between hypotheses necessarily requires that they make divergent testable predictions, and even then may not enable the relative roles of different factors to be quantified. Historical factors have doubtless had a substantial role in shaping contemporary spatial patterns of biodiversity, but deriving such *a priori* predictions and quantifying the part played by history can often prove difficult. Molecular phylogenies, with estimated dates of diversification events, provide

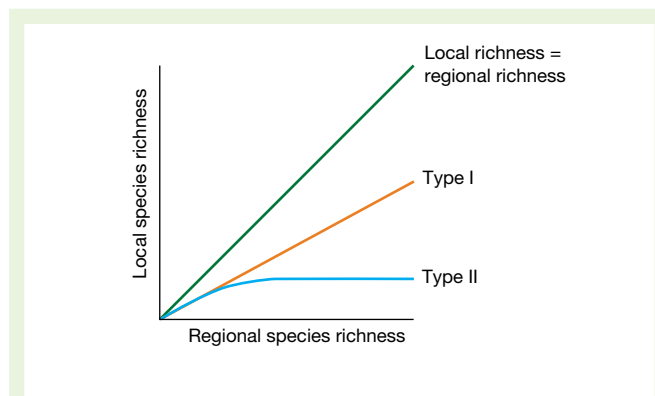


Figure 3 Relations between local and regional species richness, illustrating the form of type I and type II relationships and the limiting condition where local richness equals regional richness.

one rich source of data for testing hypotheses about the significance of history.

Relationships between local and regional richness

In exploring global variation in biodiversity, we need to understand not only the importance of differences in spatial scale for the patterns that are observed (for example, hump-shaped species–energy relationships at local scales and positive relationships at regional ones), but also how diversity at one scale might relate to that at another. Indeed, it is increasingly apparent that knowledge of the roles of pattern and process at different scales is at the very heart of an understanding of global variation in biodiversity.

Two theoretical types of relationship have been contrasted between the local richness an assemblage might attain and the species richness of the region in which that assemblage resides⁵⁴ (Fig. 3). Local richness may be directly proportional to, but less than, regional richness, following a proportional-sampling model (type I). Alternatively, as regional richness increases, local richness might attain a ceiling above which it does not rise despite continued increases in regional richness (type II).

Acknowledging a number of technical concerns^{55–57}, most real systems seem to exhibit an underlying type I relationship^{54,56,58}, not uncommonly, regional richness explains a large proportion (>75%) of variance in local richness, and local richness constitutes a marked proportion (>50%) of regional richness. For example, type I relationships have been documented for fig wasps and their parasitoids in southern and central Africa⁵⁹, tiger beetles in North America and in India⁶⁰, lacustrine fish in North America⁶¹ (Fig. 1c), and primates in Africa and in South America⁶². The predominance of type I relationships is supported by the observation that some spatial gradients in species richness are documented both for localities and regions across those gradients (with obvious implications for the interpretation of regional collations of fossil records).

A recurrent problem in studies of spatial patterns in biodiversity has been the conflation of pattern with mechanism. Nonetheless, the preponderance of examples of type I relationships, particularly where habitat type has been kept constant, backed up with other evidence (for example, the limited support for community convergence, density compensation and invasion resistance), indicates that there are not hard limits to levels of local richness⁶³. That is, local assemblages do not seem to be saturated, in the way one might have expected if ecological interactions (for example, competition, predation and parasitism) limited local richness. Three potential anomalies arise if this conclusion is correct. First, it suggests that although ecological interactions are known to be strong in some circumstances, they may typically not be sufficient to have a marked effect on species richness. Second, it may be at odds with the more-individuals

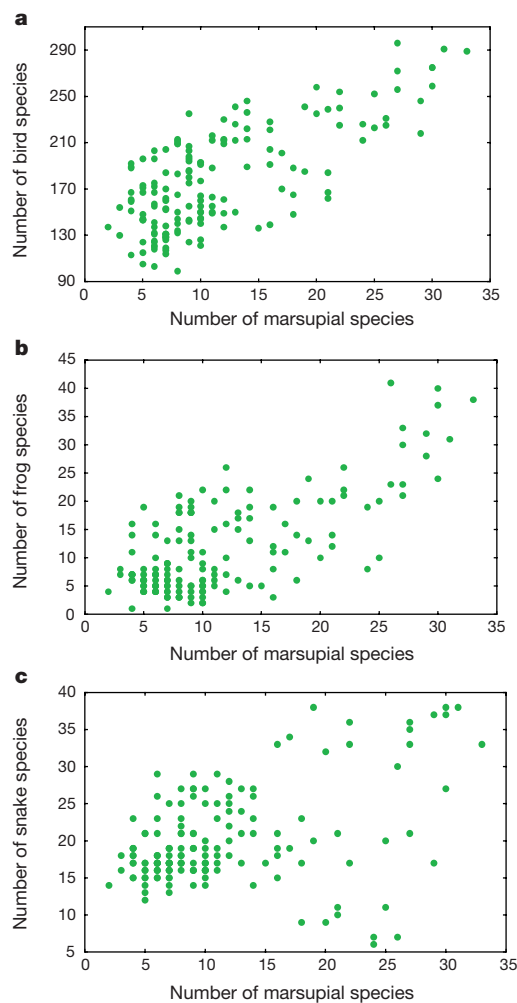


Figure 4 Relationships between species richness of different vertebrate groups (in grid cells of 240 km × 240 km) across Australia. **a**, Birds; **b**, frogs; and **c**, snakes, with marsupials (from data in ref. 79).

hypothesis as an explanation for positive species–energy relationships based on local site data, because this requires that energy levels limit the number of species that can co-occur (through limitation on the number of individuals). And third, it seems at odds with the existence of marked regional (for example, latitudinal) variation in beta diversity, although this variation may be confounded by differences in habitat composition. (Defining beta diversity as $(S/\alpha) - 1$ (where S is regional and α is local richness⁶⁴), regions having the same ratio of local-to-regional richness should have the same beta diversity³⁷.) The marked growth in studies of global patterns in biodiversity is steadily resolving such apparent anomalies by examining multiple patterns for the same assemblage, and thus generating a more coherent picture of the inter-relationships between them. But high-quality data sets documenting at high resolution the occurrence of large numbers of species over extensive areas are desperately required.

If most systems exhibit type I relationships between local and regional richness, then a prime driver of local richness seems to be regional richness. The importance of regional-scale phenomena for local-scale assemblage structure is a general one⁶⁵. A local community is assembled from a regional pool of species (this has variously been defined as the total pool of species in a region, or the pool of species in the region that is actually capable of colonizing a given site). The size and structure of this pool are influenced by regional processes, including the effects of the geophysical properties and history of

the region (its age, geology, size and climate), and broad-scale ecological or evolutionary processes, such as species migrations, invasions, speciation and regional extinction⁵⁵. They set the species composition and the abundance, body size and trophic structure of the pool from which local communities draw. Indeed, a core issue in ecology is the extent to which local assemblages can be modelled accurately as random draws from regional species pools, or conversely the extent to which local factors modify the similarity of real assemblages to randomly simulated ones⁶⁶. Almost invariably such models can explain some, often much, and occasionally most of the basic structure of local assemblages^{44,66}. Local assemblage structure and the regional context are inseparable.

Recognition of the importance of regional-scale processes and the structure of the regional species pool to local community structure has led to the emergence of macroecology, which is concerned with understanding the abundance and distribution of species at large spatial and temporal scales^{44,67}. However, although regional pools doubtless are important in structuring local assemblages, they are perhaps best seen as contributing to, rather than determining, local assemblage structure—local processes remain important. Resolving the relative contributions of local and regional processes may provide a key to understanding global patterns of biodiversity⁵⁵, but this issue once again emphasizes that patterns in biodiversity are unlikely to have a single primary cause.

Taxonomic covariance in species richness

Most major terrestrial and freshwater groups are more speciose in tropical than temperate regions, at low elevations than at high, and in forests than in deserts. One might therefore expect that the regional richness of different groups of organisms would covary positively and, because of the positive relationship between local and regional richness, local richness would do likewise. This would be important because it would simplify the development of an understanding of global patterns in biodiversity.

In practice, mismatches between the spatial occurrence of peaks in the richness of different groups have often been observed. For example, among trees, tiger beetles, amphibians, reptiles, birds and mammals, the 5% of land area across the United States and southern Canada in which the highest levels of species richness are attained overlap between some pairs of taxa, but this pattern is not a general one⁶⁸. Likewise, although the numbers of species in different, large and similar-sized areas for two groups are often significantly correlated, and may enable a very general impression of the patterns in richness of one group to be obtained from those of another, these correlations are frequently weak, of rather limited predictive value, and in some cases explained by latitudinal gradients in diversity⁶⁸ (Fig. 4). These conclusions seem to hold at finer resolutions over more constrained areas. Thus, at a scale of 10 km × 10 km squares, species-rich areas for different taxa in Britain frequently do not coincide⁶⁹. These areas are not distributed randomly, overlapping more often than expected by chance, but still at a rather low level. Likewise, different taxa are species poor or species rich in different areas of the northern region of South Africa⁷⁰.

Where positive relationships are found between the species richness of two or more groups, this may reflect patterns of sampling effort (a complication plaguing many biodiversity studies), rather than any underlying covariance. If real, then this does not necessarily imply any direct linkage between the richness of those groups. Covariance can occur because of trophic or other relations, but might also result from random effects, because groups share common determinants of richness, or even because groups differ in determinants of richness but these determinants themselves exhibit spatial covariance⁷¹.

The lack of strong positive covariance in the species richness of higher taxa is significant in that it constrains the extent to which observed patterns in biodiversity can be extrapolated from one group to another, and from exemplar groups to biodiversity at large (with implications for the planning and likely success of networks of

protected areas). The latter is particularly important given that only ~15% of the total number of species estimated to be extant has been formally described taxonomically, that the distributions of most of these remain largely unknown (a high proportion are known from only a single locality^{72,73}), and that species whose distributions are well documented are strongly biased with respect to their higher taxonomic affinities. But such outcomes are inevitable, because of the multiple forces at work in structuring global patterns of biodiversity, and because the particular outcomes observed rest fundamentally on the balance of those forces. Indeed, even where two groups exhibit similar spatial gradients in biodiversity there is substantial variation around those trends, and the details are seldom similar. In the extreme, some groups exhibit patterns of biodiversity that are entirely contrary to the norm. For example, several major taxonomic groups exhibit peaks of species richness at high or mid-latitudes (for example, aphids, sawflies, ichneumonids, braconids, bees, various groups of freshwater invertebrates, marine amphipods, and procelariiforms^{1,74}); exceptions to patterns of biodiversity tend to be observed more frequently at lower taxonomic levels than at higher levels. Which particular patterns are and are not expressed by a given taxon rest on contingencies (for example, physiology, dispersal ability, resource requirements and evolutionary history⁵⁶).

In conclusion

Development of a markedly improved understanding of the global distribution of biodiversity is one of the most significant objectives for ecologists and biogeographers. Spatial heterogeneity in species richness, in particular, is an obvious feature of the natural world. An understanding of its determinants will impinge on applied issues of major concern to humankind, including the role of biodiversity in ecosystem processes, the spread of alien invasive species, the control of diseases and their vectors, and the likely effects of global environmental change on the maintenance of biodiversity.

A substantial proportion of regional variation in species richness can be explained statistically in terms of a few environmental variables¹. This is, however, far from a predictive theory of species richness. It is the need to identify the contingencies involved in the expression of patterns in biodiversity, and to weigh their significance, that constitutes the real challenge to developing such a theory. The number of species is determined by the birth, death, immigration and emigration rates of species in an area. These rates in turn are determined by the effects of abiotic and biotic factors (the latter may be intrinsic or extrinsic to the organisms of concern) acting at local and regional scales. Although multiple factors doubtless contribute, if a factor influences biodiversity on one spatial axis (for example, latitude) then it seems reasonable to presume that all else being equal it will do so along others where the factor also varies (for example, elevation). Thus, relationships between species richness and environmental energy have been found to be associated with latitudinal, elevational and depth gradients⁷⁵. If this were the whole story, patterns in richness would seem reasonably straightforward, if not easy, to predict. However, it is not simply the current states of these factors that are important but also their historical dynamics. These have shaped variations in the distribution of different groups of organisms, in their diversification, and hence the availability of species with different attributes to exploit opportunities provided by prevailing conditions. As such, the study of global patterns in biodiversity demands insights from geneticists through to ecosystem ecologists. All concerned will need to remember that no single mechanism need adequately explain a given pattern, that observed patterns may vary with spatial scale, that processes at regional scales influence patterns observed at local ones, and that no pattern is without variations and exceptions. □

1. Gaston, K. J. & Williams, P. H. in *Biodiversity: A Biology of Numbers and Difference* (ed. Gaston, K. J.) 202–229 (Blackwell Science, Oxford, 1996).
2. Brown, J. H. & Lomolino, M. V. *Biogeography* 2nd edn (Sinauer, Sunderland, MA, 1998).
3. Stevens, G. C. The latitudinal gradient in geographical range: how so many species co-exist in the

- tropics. *Am. Nat.* **133**, 240–256 (1989).
4. Gaston, K. J. Biodiversity — latitudinal gradients. *Prog. Phys. Geogr.* **20**, 466–476 (1996).
5. Stehli, F. G., Douglas, D. G. & Newell, N. D. Generation and maintenance of gradients in taxonomic diversity. *Science* **164**, 947–949 (1969).
6. Crane, P. R. & Lidgard, S. Angiosperm diversification and paleolatitudinal gradients in Cretaceous floristic diversity. *Science* **246**, 675–678 (1989).
7. Clarke, A. & Crame, J. A. in *Marine Biodiversity: Patterns and Processes* (eds Ormond, R. F. G., Gage, J. D. & Angel, M. V.) 122–147 (Cambridge Univ. Press, Cambridge, 1997).
8. Platnick, N. I. Patterns of biodiversity: tropical vs temperate. *J. Nat. Hist.* **25**, 1083–1088 (1991).
9. Blackburn, T. M. & Gaston, K. J. Spatial patterns in the species richness of birds in the New World. *Ecography* **19**, 369–376 (1996).
10. Roy, K., Jablonski, D., Valentine, J. W. & Rosenberg, G. Marine latitudinal diversity gradients: tests of causal hypotheses. *Proc. Natl Acad. Sci. USA* **95**, 3699–3702 (1998).
11. Lyons, S. K. & Willig, M. R. A hemispheric assessment of scale dependence in latitudinal gradients of species richness. *Ecology* **80**, 2483–2491 (1999).
12. Blackburn, T. M. & Gaston, K. J. A sideways look at patterns in species richness, or why there are so few species outside the tropics. *Biodiv. Lett.* **3**, 44–53 (1996).
13. Colwell, R. K. & Hurr, G. C. Nonbiological gradients in species richness and a spurious Rapoport effect. *Am. Nat.* **144**, 570–595 (1994).
14. Terborgh, J. On the notion of favorableness in plant ecology. *Am. Nat.* **107**, 481–501 (1973).
15. Osman, R. W. & Whitlatch, R. B. Patterns of species diversity: fact or artifact? *Paleobiology* **4**, 41–54 (1978).
16. Rosenzweig, M. L. Species diversity gradients: we know more and less than we thought. *J. Mamm.* **73**, 715–730 (1992).
17. Rosenzweig, M. L. *Species Diversity in Space and Time* (Cambridge Univ. Press, Cambridge, 1995).
18. Blackburn, T. M. & Gaston, K. J. The relationship between geographic area and the latitudinal gradient in species richness in New World birds. *Evol. Ecol.* **11**, 195–204 (1997).
19. Rosenzweig, M. L. & Sandlin, E. A. Species diversity and latitudes: listening to area's signal. *Oikos* **80**, 172–176 (1997).
20. MacArthur, R. H. & Connell, J. H. *The Biology of Populations* (Wiley, New York, 1966).
21. Gaston, K. J. & Blackburn, T. M. A critique for macroecology. *Oikos* **84**, 353–368 (1999).
22. Lawton, J. H. Patterns in ecology. *Oikos* **75**, 145–147 (1996).
23. Rosenzweig, M. L. & Abramsky, Z. in *Species Diversity in Ecological Communities* (eds Ricklefs, R. E. & Schluter, D.) 52–65 (Univ. Chicago Press, Chicago, 1993).
24. Currie, D. J. & Paquin, V. Large-scale biogeographical patterns of species richness of trees. *Nature* **329**, 326–327 (1987).
25. Turner, J. R. G., Gatehouse, C. M. & Corey, C. A. Does solar energy control organic diversity? Butterflies, moths and the British climate. *Oikos* **48**, 195–205 (1987).
26. Turner, J. R. G., Lennon, J. J. & Lawrenson, J. A. British bird species distributions and energy theory. *Nature* **335**, 539–541 (1988).
27. Adams, J. M. & Woodward, F. I. Patterns in tree species richness as a test of the glacial extinction hypothesis. *Nature* **339**, 699–701 (1989).
28. Currie, D. J. Energy and large-scale patterns of animal- and plant-species richness. *Am. Nat.* **137**, 27–49 (1991).
29. Wright, D. H., Currie, D. J. & Maurer, B. A. in *Species Diversity in Ecological Communities* (eds Ricklefs, R. E. & Schluter, D.) 66–74 (Univ. Chicago Press, Chicago, 1993).
30. Kerr, J. T., Vincent, R. & Currie, D. J. Lepidopteran richness patterns in North America. *Ecoscience* **5**, 448–453 (1998).
31. Kerr, J. T. & Packer, L. The environmental basis of North American species richness patterns among *Epicauda* (Coleoptera: Meloidae). *Biodiv. Conserv.* **8**, 617–628 (1999).
32. Rutherford, S., D'Hondt, S. & Prell, W. Environmental controls on the geographic distribution of zooplankton diversity. *Nature* **400**, 749–753 (1999).
33. Lennon, J. J., Greenwood, J. J. D. & Turner, J. R. G. Bird diversity and environmental gradients in Britain: a test of the species-energy hypothesis. *J. Anim. Ecol.* (in the press).
34. Kerr, J. T. & Currie, D. J. The relative importance of evolutionary and environmental controls on broad-scale patterns of species richness in North America. *Ecoscience* **6**, 329–337 (1999).
35. Tilman, D. & Pacala, S. in *Species Diversity in Ecological Communities* (eds Ricklefs, R. E. & Schluter, D.) 13–25 (Univ. Chicago Press, Chicago, 1993).
36. Gaston, K. J. & Blackburn, T. M. Mapping biodiversity using surrogates for species richness: macro-scales and New World birds. *Proc. R. Soc. Lond. B* **262**, 335–341 (1995).
37. Chown, S. L. & Gaston, K. J. Patterns in procelariiform diversity as a test of species-energy theory in marine systems. *Evol. Ecol. Res.* **1**, 365–373 (1999).
38. Martin, T. E. Species-area slopes and coefficients: a caution on their interpretation. *Am. Nat.* **118**, 823–837 (1981).
39. Palmer, M. W. & White, P. S. Scale dependence and the species-area relationship. *Am. Nat.* **144**, 717–740 (1994).
40. Latham, R. E. & Ricklefs, R. E. Global patterns of tree species richness in moist forests: energy-diversity theory does not account for variation in species richness. *Oikos* **67**, 325–333 (1993).
41. Francis, A. P. & Currie, D. J. Global patterns of tree species richness in moist forests: another look. *Oikos* **81**, 598–602 (1998).
42. Ricklefs, R. E., Latham, R. E. & Qian, H. Global patterns of tree species richness in moist forests: distinguishing ecological influences and historical contingency. *Oikos* **86**, 369–373 (1999).
43. Cousins, S. H. Species richness and the energy theory. *Nature* **340**, 350–351 (1989).
44. Gaston, K. J. & Blackburn, T. M. *Pattern and Process in Macroecology* (Blackwell Science, Oxford, in the press).
45. Srivastava, D. S. & Lawton, J. H. Why more productive sites have more species: an experimental test of theory using tree-hole communities. *Am. Nat.* **152**, 510–529 (1998).
46. Gaston, K. J., Blackburn, T. M. & Lawton, J. H. Interspecific abundance-range size relationships: an appraisal of mechanisms. *J. Anim. Ecol.* **66**, 579–601 (1997).
47. Wright, D. H. Species-energy theory: an extension of species area-theory. *Oikos* **41**, 496–506 (1983).
48. Hayden, B. P. Ecosystem feedbacks on climate at the landscape scale. *Phil. Trans. R. Soc. Lond. B* **353**, 5–18 (1998).
49. Zheng, X. Y. & Eltahir, E. A. B. The role of vegetation in the dynamics of West African monsoons. *J. Clim.* **11**, 2078–2096 (1998).
50. Collar, N. J., Crosby, M. J. & Stattersfield, A. J. *Birds to Watch 2: The World List of Threatened Birds* (BirdLife International, Cambridge, 1994).
51. Schläpfer, F. & Schmid, B. Ecosystem effects of biodiversity: a classification of hypotheses and

- exploration of empirical results. *Ecol. Appl.* **9**, 893–912 (1999).
52. Tilman, D. The ecological consequences of changes in biodiversity: a search for general principles. *Ecology* **80**, 1455–1474 (1999).
53. Holmes, R. T. & Sturges, F. W. Avian community dynamics and energetics in a northern hardwoods ecosystem. *J. Anim. Ecol.* **44**, 175–200 (1975).
54. Cornell, H. V. & Lawton, J. H. Species interactions, local and regional processes, and limits to the richness of ecological communities: a theoretical perspective. *J. Anim. Ecol.* **61**, 1–12 (1992).
55. Huston, M. A. Local processes and regional patterns: appropriate scales for understanding variation in the diversity of plants and animals. *Oikos* **86**, 393–401 (1999).
56. Lawton, J. H. Are there general laws in ecology? *Oikos* **84**, 177–192 (1999).
57. Srivastava, D. S. Using local-regional richness plots to test for species saturation: pitfalls and potentials. *J. Anim. Ecol.* **68**, 1–16 (1999).
58. Caley, M. J. & Schluter, D. The relationship between local and regional diversity. *Ecology* **78**, 70–80 (1997).
59. Hawkins, B. A. & Compton, S. G. African fig wasp communities: undersaturation and latitudinal gradients in species richness. *J. Anim. Ecol.* **61**, 361–372 (1992).
60. Pearson, D. L. & Juliano, S. A. in *Species Diversity in Ecological Communities* (eds Ricklefs, R. E. & Schluter, D.) 194–202 (Univ. Chicago Press, Chicago, 1993).
61. Griffiths, D. Local and regional species richness in North American lacustrine fish. *J. Anim. Ecol.* **66**, 49–56 (1997).
62. Eeley, H. A. C. & Lawes, M. J. in *Primate Communities* (eds Feagle, J. G., Janson, C. & Reed, K. E.) 191–219 (Cambridge Univ. Press, Cambridge, 1999).
63. Cornell, H. V. Unsaturating and regional influences on species richness in ecological communities: a review of the evidence. *Ecoscience* **6**, 303–315 (1999).
64. Whittaker, R. H. Vegetation of the Siskiyou mountains, Oregon and California. *Ecol. Monogr.* **30**, 279–338 (1960).
65. Ricklefs, R. E. & Schluter, D. (eds) *Species Diversity in Ecological Communities* (Univ. Chicago Press, Chicago, 1993).
66. Gotelli, N. J. & Graves, G. R. *Null Models in Ecology* (Smithsonian Institution Press, Washington DC, 1996).
67. Brown, J. H. *Macroecology* (Univ. Chicago Press, Chicago, 1995).
68. Flather, C. H., Wilson, K. R., Dean, D. J. & McComb, W. C. Identifying gaps in conservation networks: of indicators and uncertainty in geographic-based analyses. *Ecol. Appl.* **7**, 531–542 (1997).
69. Prendergast, J. R., Quinn, R. M., Lawton, J. H., Eversham, B. C. & Gibbons, D. W. Rare species, the coincidence of diversity hotspots and conservation strategies. *Nature* **365**, 335–337 (1993).
70. van Jaarsveld, A. S. *et al.* Biodiversity assessment and conservation strategies. *Science* **279**, 2106–2108 (1998).
71. Gaston, K. J. Biodiversity — congruence. *Prog. Phys. Geogr.* **20**, 105–112 (1996).
72. Stork, N. E. in *Biodiversity II: Understanding and Protecting our Biological Resources* (eds Reaka-Kudla, M. L., Wilson, D. E. & Wilson, E. O.) 41–68 (Henry, Washington DC, 1997).
73. Andersen, M., Thornhill, A. & Koopowitz, H. in *Tropical Forest Remnants: Ecology, Management, and Conservation of Fragmented Communities* (eds Laurance, W. F. & Bierregaard, R. O. Jr) 281–291 (Univ. Chicago Press, Chicago, 1997).
74. Chown, S. L., Gaston, K. J. & Williams, P. H. Global patterns in species richness of pelagic seabirds: the Procellariiformes. *Ecography* **21**, 342–350 (1998).
75. Rohde, K. Latitudinal gradients in species diversity: the search for the primary cause. *Oikos* **65**, 514–527 (1992).
76. Judas, M. The species-area relationship of European Lumbricidae (Annelida, Oligochaeta). *Oecologia* **76**, 579–587 (1988).
77. Patterson, B. D., Stotz, D. F., Solari, S., Fitzpatrick, J. W. & Pacheco, V. Contrasting patterns of elevational zonation for birds and mammals in the Andes of southeastern Peru. *J. Biogeogr.* **25**, 593–607 (1998).
78. O'Brien, E. M. Climatic gradients in woody plant species richness: towards an explanation based on an analysis of southern Africa's woody flora. *J. Biogeogr.* **20**, 181–198 (1993).
79. Pianka, E. R. & Schall, J. J. in *Ecological Biogeography of Australia* Vol. 3 (ed. Keast, A.) 1677–1694 (Junk, The Hague, 1981).

Acknowledgements

I thank T. M. Blackburn, S. L. Chown, A. Clarke, S. Gaston, P. H. Warren, T. J. Webb and F. I. Woodward for generous discussion and comments, and J. J. D. Greenwood and the British Trust for Ornithology, D. Griffiths, J. T. Kerr, J. J. Lennon, E. M. O'Brien, B. D. Patterson and R. J. Whittaker for kindly providing data. K.J.G. is a Royal Society University Research Fellow.

The diversity–stability debate

Kevin Shear McCann

1205 Docteur Penfield Avenue, Department of Biology, McGill University, Montreal, Quebec, Canada H3A 1B1

There exists little doubt that the Earth's biodiversity is declining. The Nature Conservancy, for example, has documented that one-third of the plant and animal species in the United States are now at risk of extinction. The problem is a monumental one, and forces us to consider in depth how we expect ecosystems, which ultimately are our life-support systems, to respond to reductions in diversity. This issue — commonly referred to as the diversity–stability debate — is the subject of this review, which synthesizes historical ideas with recent advances. Both theory and empirical evidence agree that we should expect declines in diversity to accelerate the simplification of ecological communities.

We now realize that the world's flora and fauna are disappearing at rates greater than the mass extinction events whose collapses punctuate the fossil record^{1–3}. It is also true that species invasions have been elevated to unprecedented rates accompanying the increased globalization of our world^{4,5}. These high rates of extinction and invasion put ecosystems under enormous stress, making it critical that we understand how the loss, or addition, of a species influences the stability and function of the ecosystems we rely on. We are, in a very real sense, deconstructing the Earth under the implicit assumption that ecosystems have evolved the ability to withstand such assault without collapse.

Several advances in the diversity–stability debate form a conceptual thread that suggests that diversity can be expected, on average, to give rise to ecosystem stability. This does not infer that diversity is the driver of this relationship. Instead, diversity can be regarded as the passive recipient of important ecological mechanisms that are inherent in ecosystems. One promising mechanism that has been proposed recently is that weakly interacting species stabilize community dynamics by dampening strong, potentially destabilizing consumer–resource interactions⁶. Empirical descriptions of the distribution of interaction strengths in real communities are consistent with this theory. If this is true then, all else being equal, decreasing biodiversity will be accompanied by increases in average interaction strengths within ecosystems, and a concomitant decrease in ecosystem stability.

Historical perspectives of the diversity–stability debate

The relationship between diversity and stability has fascinated ecologists. Before the 1970s, ecologists believed that more diverse communities enhanced ecosystem stability^{7–9}. A strong proponent of this view was Charles Elton⁸, who argued that “simple communities were more easily upset than that of richer ones; that is, more subject to destructive oscillations in populations, and more vulnerable to invasions”. In fact, both Odum⁷ and Elton⁸ arrived at similar conclusions based on the repeated observation that greatly simplified terrestrial communities are characterized by more violent fluctuations in population density than diverse terrestrial communities. For example, invasions most frequently occur on cultivated land where human influence had produced greatly simplified ecological communities, and outbreaks of phytophagous insects occur readily in boreal forests but are unheard of in diverse tropical forests. These observations led Elton⁸ to believe that

complex communities, constructed from many predators and parasites, prevented populations from undergoing explosive growth. His ideas were closely akin to MacArthur⁹, who reasoned that multiplicity in the number of prey and predator species associated with a population freed that population from dramatic changes in abundance when one of the prey or predator species declined in density.

These early intuitive ideas were challenged by the work of Robert May¹⁰ in 1973. May turned to mathematics to rigorously explore the diversity–stability relationship. By using linear stability analysis on models constructed from a statistical universe (that is, randomly constructed communities with randomly assigned interaction strengths), May found that diversity tends to destabilize community dynamics. Other ecologists, using similar approaches, found results that were consistent with this hypothesis^{11,12}. The results were puzzling, as real ecosystems were undoubtedly complex and diverse. The results also seemed to counter the ideas of Elton⁷, Odum⁸ and MacArthur⁹. Yodzis¹³ heightened this paradox by showing that models structured from compiled food-web relationships, with plausible interaction strengths, were generally more stable than randomly constructed food webs. Although the early food-web data that Yodzis structured his models around were incomplete, these data reflected real feeding relationships. Yodzis' result indicated that interaction strength was probably crucial to stability; but the exact reason for this remained elusive. If diversity and stability were positively correlated, as early empirical evidence had indicated, then more had to be happening than simply increasing the number of species and the number of pathways. Something fundamental was missing from the early arguments.

In the remainder of this paper I review recent investigations of the diversity–stability debate. I first discuss a change in perspective that is beginning to allow us to unravel this long-standing problem and then review two different lines of investigation. One approach has searched for a general diversity–stability relationship, and a second, more mechanistic approach has sought a relationship between food-web structure and stability.

Changing perspectives

Much of ecological theory is based on the underlying assumption of equilibrium population dynamics. Although this assumption is aesthetically pleasing, in that it suggests the balance of nature is infinitely precise, an alternative and viable ecological perspective exists. As real populations are variable, it is possible that the persistence of complex communities depends to some degree on population fluxes (that is, the



Figure 1 The Ecotron experiment creates model multitrophic community assemblages containing plants, herbivores, parasitoids and decomposers in 16 different chambers. The Ecotron is an ambitious attempt to bridge the scale between field communities and laboratory experiments. (Photographs show the inside of an Ecotron chamber and a technical service corridor between two banks of chambers; courtesy of the Centre for Population Biology, Imperial College at Silwood Park.)

fairly regular waxing and waning of a population's density). Such background population variability, whether driven by biotic or abiotic processes, can provide species with the opportunity to respond differentially to their environment. In turn, these differential species responses weaken the destructive potential of competitive exclusion.

Because such variability can significantly change our understanding of ecological interactions^{6,14–19}, ecologists have begun to relax equilibrium-based measures of stability. A recent theoretical analysis¹⁷ has shown that population fluctuations, driven by competition, can actually promote the persistence of large numbers of competing phytoplankton communities on a minimal number of limiting resources (but greater than two resources). Coexistence was found to rely on the fluctuation in population densities, while community-level densities (the summation of the competing plankton densities) varied little. We will see that a similar relationship appears in diversity–stability experiments. Here, too, the evidence points to variable population densities that sum to produce a relatively constant biomass at the community level.

Definitions of stability

Definitions of stability in ecology can be classified generally into two categories (Table 1) — stability definitions that are based on a system's dynamic stability, and stability definitions that are based on a system's ability to defy change (resilience and resistance in Table 1). Despite the breadth of definitions, ecological theory has tended traditionally to rely on the assumption that a system is stable if, and only if, it is governed by stable equilibrium dynamics (that is, equilibrium stability and equilibrium resilience). As discussed in the previous section, these are strong assumptions with no *a priori* justification. In fact, the variable nature of population dynamics found both in field and in laboratory experiments has led experimentalists to use measures of variability as indices of a system's stability. This discontinuity between stability experiments and equilibrium-based theory has made it difficult to unite theory and experiment in the diversity–stability debate.

More general definitions of stability exist. In Table 1, general stability is defined such that stability increases as population

densities move further away from extremely low or high densities. This is a broad definition, including equilibrium and non-equilibrium dynamics as well as subsuming the definition of permanence¹⁸ (a population is considered permanent if the lower limit to its density is greater than zero). Because the definition of general stability implies decreased variability (owing to greater limits on density), it is closely related to field measurements of stability, which tend to rely on variability in population or community densities as a measure of stability. One can also extend equilibrium resilience to a less biologically restrictive form by defining resilience as the return time after a perturbation to an equilibrium or a non-equilibrium attractor (Table 1). In a nonlinear system there is no reason to believe that an equilibrium that attracts weakly in a local setting (near the equilibrium) also attracts weakly far away from the equilibrium, where the issue of a species' permanence is resolved¹⁸. For the remainder of the paper, unless stated otherwise, the definitions of general stability and variability will be used to consider empirical and theoretical results on the relationship between diversity and stability under a common framework.

The search for a general diversity–stability relationship

In 1982, David Tilman began a long-term study to delineate experimentally the relationship between diversity and stability in plant communities. The undertaking involved four grassland fields at Cedar Creek Natural History Area, Minnesota, divided into over 200 experimental plots, and gathered information on species richness, community biomass and population biomass through time. The results of this and other extensive studies converge on the finding that diversity within an ecosystem tends to be correlated positively with plant community stability (that is, decreased coefficient of variability in community biomass)^{20–23}. At the same time, diversity seems to show little influence on population variability²². The basic arguments for a positive relationship between diversity and stability for primary producers at the community level have been classified into two, not mutually exclusive, hypotheses called the averaging effect²⁴ and the negative covariance effect²⁵ (see Table 2 for the underlying logic behind these ideas). In essence, these hypotheses argue that diversity

(species richness) increases stability at the community level because diverse plant communities respond differentially to variable background processes. The differential responses of populations sum, through time, to give stable community dynamics.

If diversity and stability are positively correlated, then both the averaging and negative covariance effect predict that population variance has to scale as a function of mean population densities in a precise way (see Table 2). Tilman²⁰ has used these predictions to show that his field experiments are consistent with the interpretation that increasing diversity increases community stability. Although this is a clever combination of theory and experiment, it cannot be used to infer that diversity is responsible directly for stability^{26,27}. As a counter example, no correlation was found between diversity and stability at the cross-ecosystem scale²⁶. Other experiments have found that the positive diversity–stability correlation is not a pure species effect (that is, a diversity effect), and have indicated that ecosystem function and stability are more directly related to functional diversity (for example, graminoids or grasses, nitrogen-fixing legumes and other herbs)^{27–30}.

In a similar manner, plant community stability and productivity in European grasslands were shown to be tightly coupled to the functional diversity of mutualistic arbuscular mycorrhizal fungi (AMF)³¹. In this system, large fluctuations in plant biomass were associated with low diversities of AMF, whereas more constant biomass and greater productivity accompanied high AMF diversities. This study highlights that higher-level interactions, which are inherent in food webs (for example, microbial interactions, herbivory and predation), are of great importance in understanding the relationship between the diversity and stability of whole ecological communities. The complexity of whole ecological communities — the basis from which Odum, Elton and MacArthur formed their diversity–stability hypotheses — cannot manifest itself in experiments that focus on single trophic levels.

Field tests at the scale of the food web are few in number. But one thorough examination³² tested seven different stability–diversity criteria in the grazing ecosystem of the Serengeti under naturally variable conditions (that is, strong seasonal changes). Of these seven stability measures, five were positively related to diversity whereas two were unrelated to diversity. The study found that greater diversity reduced the relative magnitudes of fluctuations in productivity induced by seasonal change. Although a relationship between stability and diversity exists within the Serengeti, the evidence again points to the importance of functional species in understanding this relationship. For example, the grazing-tolerant plant species have a disproportionately large role in the Serengeti community dynamics by preventing herbivores from dramatically reducing plant biomass.

The paucity of field tests at the scale of the food web reflects the fact that such experiments require an enormous undertaking. As an alternative, ecologists have approached this problem by investigating how diversity influences stability and function within a multitrophic setting in controlled microcosm experiments (often referred to as bottle experiments as they are attempts to create realistic ecological

Table 1 Definitions of stability

Term	Definition
Definitions of dynamic stability	
Equilibrium stability	A discrete measure that considers a system stable if it returns to its equilibrium after a small perturbation away from the equilibrium. A stable system, therefore, has no variability in the absence of perturbations.
General stability	A measure which assumes that stability increases as the lower limit of population density moves further away from zero. Under non-equilibrium dynamics, such limits to population dynamics generally imply a decrease in population variance (see variability definition below).
Variability	The variance in population densities over time, usually measured as the coefficient in variation. Common in experimental tests of stability.
Definitions of resilience and resistance stability	
Equilibrium resilience	A measure of stability that assumes system stability increases as time required to return to equilibrium decreases after a perturbation. A rapid response means that a system recoils rapidly back to its equilibrium state.
General resilience	A measure of stability that assumes system stability increases as return time to the equilibrium/non-equilibrium solution decreases after a perturbation. A rapid response means that a system recoils rapidly back to its equilibrium/non-equilibrium state.
Resistance	A measure of the degree to which a variable changes after a perturbation. Frequently used as a discrete measure that assesses a community's ability to resist invasion (that is, if an invader fails, the community resists invasion).

communities within a controlled setting). The main advantage of microcosms is that the experiments can easily be manipulated and replicated³³. Nonetheless, the issue of how scale influences outcome looms over microcosm experiments — can we extrapolate results to the whole ecosystem? Ambitious experimental set ups such as that currently underway in the Ecotron (Fig. 1) are attempting to bridge the gap between the complexity of real field communities and the simplicity of laboratory or greenhouse experiments.

The evidence that has emerged from microcosm experiments, regardless of scale and system type (that is, terrestrial or aquatic), has tended to agree that diversity is positively related to ecosystem stability^{34–39}. In addition, and consistent with field experiments on plant communities, experiments using aquatic microcosms have shown that population-level variation is relatively uninfluenced by diversity, whereas community-level variance tends to decrease with increased diversity³⁵. Two ideas have been advanced in explanation of these findings. One explanation is that increasing diversity increases the odds that at least some species will respond differentially to variable conditions and perturbations^{37–39}. The second is that greater diversity increases the odds that an ecosystem has functional redundancy by containing species that are capable of functionally replacing important species^{37–39}. Taken together, these two notions have been called the insurance hypothesis (Table 2). This idea has been extended to suggest that the greater the variance in species' responses contained in a community then the lower the species richness required to insure the ecosystem⁴⁰. As with the averaging and negative covariance effect,

Table 2 General diversity–stability theory

Theory	Underlying logic
Averaging effect ^{24,25}	Assume covariances between species are zero and variance (s_i^2) in abundance of individual species i in a plant community is equal to cm_i^2 , where c and z are constants and m_i is the mean density of species i . Given that all k species of a community are equal in abundance and sum to m (that is, $m_i = m/k$), then the coefficient of variation (CV) of community abundance can be determined as: $CV = 100s/m = 100(c/k)^{1/2}$ For the case $z > 1$, increasing k (species number) decreases the variation in biomass for the plant community.
Negative-covariance effect ²⁵	If covariances between species (say, species a and b) are negative (that is, $cov(a,b) < 0$), then the variance in the abundance of two species $s_{(a+b)}^2 = s_a^2 + s_b^2 + 2cov(a,b)$ will be less than the sum of the individual variances (that is, $s_a^2 + s_b^2$), and so will decrease overall biomass variance in the plant community.
Insurance effect ^{37–40}	An ecosystem's ability to buffer perturbations, loss in species and species invasions is dependent on the redundancy of the species having important stabilizing roles, as well as on the ability of the species in the community to respond differentially to perturbations. Increasing diversity increases the odds that such species exist in an ecosystem. This idea has been extended ⁴⁰ to suggest that the greater the variance of species' responses in a community then the lower the species richness required to buffer an ecosystem.

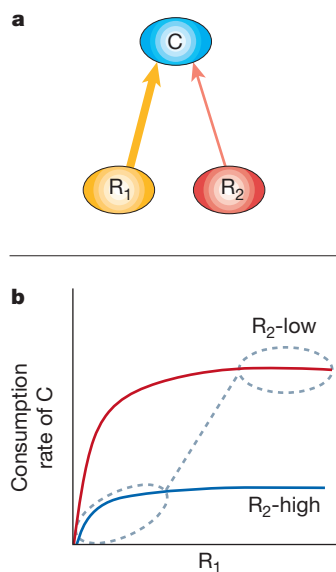


Figure 2 Consumer–resource interactions. **a**, A simple food-web diagram depicting a strong consumer–resource interaction ($C-R_1$) coupled to a weak consumer–resource interaction ($C-R_2$). **b**, Consumption rates by consumer C of R_1 for two different densities of R_2 . Because the resources negatively covary, the actual consumption response is qualitatively similar to a combination of these two curves (dashed circles and line). In the presence of R_2 , resource R_1 is less influenced by consumption when at low densities.

which are intimately related, the insurance hypothesis does not infer that diversity actively promotes stability.

In summary, the results indicate that within an ecosystem, diversity tends to be correlated positively with ecosystem stability. This correlation does not necessarily extend to population-level stability. Much work is still required to determine the driver of the positive diversity–stability relationship; however, it seems that community-level stability is dependent on the differential response of species or functional groups to variable conditions, as well as the functional redundancy of species that have important stabilizing roles. I now turn to a separate approach that has not focused on diversity, but has concentrated instead on understanding the implications of common food-web structures on stability.

Food-web structure and stability

In an important theoretical contribution, Chesson and Huntley⁴¹ showed that diversity cannot be maintained by variation alone. Rather, persistence of diversity requires the two following components: the existence of flux or variability in ecosystems; and populations capable of differentially exploiting this flux or variability. Regardless of the source of the variability (for example, whether spatially or temporally generated), their results indicate that coexistence requires that populations must be released, either directly or indirectly, from the limiting influences of species interactions such as predation and competition. Species interactions, therefore, must be important in maintaining and promoting persistence in diverse communities in spite of, and perhaps because of, the variability that underlies ecosystems. Several more specific models can be included under this general framework, and all reveal that flux interacting with specific biotic, nonlinear responses can promote persistence^{6,14,17}. We now turn to a set of food-web models that have shown how persistent, complex ecosystems can be an outcome of this combination of flux and density-dependent food-web interactions (that is, competitive and predatory influences that vary with density).

The weak-interaction effect

Over the past decade, ecologists have begun to replace the conceptu-

alization of the ecosystem as a linear food chain with the view that food webs are highly interconnected assemblages^{42–45} characterized by recurrent food-web structures (for example, omnivory and apparent competition). Because combinations of competition and predation can represent these common food-web structures, the use of simple food-web modules has been advocated⁴⁶ to explore the repercussions of these ubiquitous species interactions.

Several model investigations have grown out of this approach to show that natural food-web structures can, indeed, enhance ecosystem stability^{6,47–49}. These food-web models are extensions of a bioenergetic consumer–resource model⁵⁰ that constrains parameters to empirically determined relationships of body size. The approach is akin to the dynamic modelling of a population's energy budgets through time, and has the important consequence of placing food-web models within a biological universe with reasonable constraints operating on energy flow between any consumer–resource interaction—a feature that is considerably different from a statistical universe. The result is that increasing diversity can increase food-web stability under one condition: the distribution of consumer–resource interaction strengths must be skewed towards weak interaction strengths. I will refer to this as the weak-interaction effect (Table 3), and to connect this to general diversity–stability theory I briefly discuss the stabilizing mechanisms behind this effect.

Two general stabilizing mechanisms underlie the weak-interaction effect. First, the weak-interaction effect generates negative covariances and promotes community-level stability. Second, these negative covariances ensure that the weak interactors dampen the destabilizing potential of strong consumer–resource interactions. These mechanisms can be best understood with a simple example.

Figure 2a depicts a simple food-web interaction in which a strong consumer–resource interaction ($C-R_1$) is coupled to a weak consumer–resource interaction ($C-R_2$). Being a weakly interacting species, R_2 is an inferior competitor whose ability to persist is mediated by the top predator. This food-web relationship ensures that the resources negatively covary. For example, R_2 is released from competitive limitation to flourish whenever R_1 is suppressed by high densities of consumer C . This occurs because R_2 is weakly coupled to C and so is not strongly influenced by high densities of C . In this manner, the weak interaction drives the differential responses of species.

We can use the knowledge of this negative covariance to determine qualitatively the consumption rate of C on its preferred resource, R_1 . Figure 2b depicts C 's consumption rate on R_1 under two different densities of R_2 , assuming an optimally foraging, type II multispecies functional response (for full details, see refs 49, 51). High densities of R_2 reduce the overall consumption rates on R_1 . Because the resources negatively covary, then for low densities of R_1 we expect C 's consumption rates to fall on the R_2 -high curve in Fig. 2 (the lower dashed circle). Similarly, for high densities of R_1 we expect consumption rates to be on the R_2 -low curve in Fig. 2 (the upper dashed circle). Piecing these functions together we see that the asynchrony in resource densities drives a sigmoid-shaped response that is qualitatively similar to what ecologists refer to as a type III functional response. This has the non-equilibrium effect of releasing the prey (R_1) from strong consumptive pressures when it is at low densities, and thereby the weak interaction dampens the oscillatory potential of the strong $C-R_1$ interaction. Consistent with the above discussion is the fact that investigators have found that donor control (in which a consumer responds numerically to a resource but has no influence on resource dynamics) also promotes community stability^{45,48}. Donor control can generate differential responses of species by allowing species using these resources to disconnect themselves from fluxes that are inherent to the community.

I have discussed the weak-interaction effect within the context of relatively simple food-web modules. Does the effect operate for real, complex communities with enormous numbers of direct and indirect interactions? It is still too early to tell, but Kokkoris *et al.*⁵² followed the distribution of interaction strengths as competitive

model communities were assembled. They found that as the assembly process progressed, larger permanent communities (that is, with a lower limit above zero) attained lower mean interaction strengths. They also found that communities with lower mean interaction strength were more resistant to invasion. These results are encouraging and indicate that the weak-interaction effect might scale to the whole ecosystem. If the weak-interaction mechanism is operating in real communities then the distributions of interaction strengths will be skewed towards weak interactions in order that a few potentially excitable consumer–resource interactions are muted. I now turn to empirical investigations of food-web structure and stability, first reviewing experiments concerned with the distribution of interaction strengths in natural communities before examining experiments that have investigated directly the influence of food-web structure on stability.

Interaction strength and species invasions

Although quantitative field estimates of interaction strength are still in the process of development, work by a few ecologists has enabled a preliminary glimpse into the nature of the distributions of interaction strength within real food webs^{53–57}. The early data indicate unequivocally that distributions of interaction strength are strongly skewed towards weak interactions^{53–57}. Nonetheless, these experiments also highlight that the removal or addition of a single key species can have pronounced impacts on the dynamics and persistence of the species in the enclosure or enclosure. For example, experimental removal of the predatory starfish, *Pisaster ochraceus*, resulted in greatly simplified lower-intertidal communities because the mussel, *Mytilus californianus*, competitively dominates all other sessile benthic organisms when freed from predation⁵⁸.

A recent experiment⁵⁹ has confirmed the abundance of weak interactions in ecosystems, but showed that weak average interaction strength in a rocky intertidal community tends to be correlated with high variability in interaction strength. In this study, variation in the magnitude of the weak interactions seemed to excite spatial variation in community structure. The variation in interaction strength may be important in generating landscape-scale variation that promotes the maintenance of diversity, an area that demands further investigation.

It is important to know if these phenomena can be extended beyond the scale of the enclosure/exclosure experiment. Species invasions may be seen as the uncontrolled version of species addition experiments. Similar to the experiments described above, the current evidence indicates that, although most species invasions have a weak impact on ecosystems⁶⁰, the occasional invasive species alters an ecosystem profoundly^{61–63}. For example, a recent study⁶¹ used stable isotopes to document energy flow through food webs. Lakes that were uninvaded by bass were compared with lakes that had just been invaded, and the recently invaded lakes showed marked differences in energy flow patterns (implying a severely altered food-web structure) as well as rapid declines in forage fish diversity⁶¹. These results indicate that the addition of a single species can precipitate a form of ecosystem collapse that sends a wave of extinction through the ecosystem. Another noteworthy case concerns the introduction of the large predatory fish, the Nile perch (*Lates niloticus*), in Lake Victoria in the 1950s. The addition of the Nile perch was followed by a sequence of amazing ecological and genetic changes that culminated in a cascade of cichlid extinctions⁶³. Overall, however, the invasion literature is harmonious with enclosure/exclosure experiments⁶⁰ — most invasions have a weak impact with infrequent occurrences of an invasive species capable of precipitating monumental changes to an ecosystem.

Food-web structure and stability experiments

Some direct experimental tests of stability and food-web structure exist. In a clever experimental manipulation, Fagan⁶⁴ tested community response to a perturbation (aphicide application) as a function of the degree of omnivory. Fagan accomplished this by controlling

Table 3 Food-web structure and stability theory

Theory	Underlying logic
Weak-interaction effect ^{6,46–48}	Weak interactions serve to limit energy flow in a potentially strong consumer–resource interaction and, therefore, to inhibit runaway consumption that destabilizes the dynamics of food webs. In addition, the weak interactions serve to generate negative covariances between resources that enable a stabilizing effect at the population and community level. The negative covariances ensure that consumers have weak consumptive influences on a resource when the resource is at low densities. See text and Fig. 2 for further clarification.
	Berlow ⁶⁷ suggested an additional influence of weak interactions. Weak interactions in intertidal communities seem to be extremely variable in strength, and as a result may drive spatial variability in community structure. This community variability in space can provide a canvas for species to respond differentially, and so may further promote the maintenance of diversity.

the relative proportion of nonomnivorous damselbugs versus omnivorous wolf spiders in arthropod assemblages of the Mount Saint Helen ‘blowdown zone’. The results showed that increasing the degree of community omnivory (that is, increasing the proportion of wolf spiders) decreased variation in the population responses after an aphicide application.

In an earlier investigation, de Ruiter *et al.*⁶⁵ investigated model communities constructed from empirical estimates based on some well-studied food webs from native and agricultural soils. Their results are consistent with the experiments on interaction strength and the weak-interaction effect discussed above, as most interactions had only a negligible impact on community dynamics. Although their results indicate that energetics are important in constraining interaction strength, they found no positive correlation between feeding rates and community impact. This, too, is consistent with the weak-interaction effect. Similarly, experiments on both terrestrial and aquatic microcosms have tended to find that increasing the number of prey items enhances stability^{66–68}, although one microcosm experiment⁶⁹ found that the addition of an alternate prey destabilized community dynamics. This last case can be reconciled with other experiments as the alternate prey introduced was efficient fare for the predator. In essence, the alternative prey energetically fuelled the predator, and so the experiment may be viewed as evidence that a strong consumer–resource interaction is potentially destabilizing.

Conclusion

Taken together, recent advances indicate that diversity can be expected, on average, to give rise to ecosystem stability. The evidence also indicates that diversity is not the driver of this relationship; rather, ecosystem stability depends on the ability for communities to contain species, or functional groups, that are capable of differential response. All of these views are consistent with the ideas put forth by the influential figures of Odum⁷, Elton⁸, MacArthur⁹ and May¹⁰. May’s result reflects the fact that random distributions (that is, a null universe where dynamics are influenced by diversity alone), on average, do not create the necessary tension between community members that forces differential response and community stability. In a randomly constructed community, for example, strong interactions are not necessarily coupled to weak interactions that mute their destabilizing potential. In fact, one can expect that random communities will generally not create such couplings, and so tend to produce diverse communities with complex, wildly oscillatory dynamics. Furthermore, Odum, Elton and MacArthur recognized that real food webs contain a complex array of energetic pathways that can act as buffers against dramatic population explosions. Specifically, MacArthur’s hypothesis — that greater connectance drives community and ecosystem stability — seems a strong possibility provided most pathways are constructed from weak interactions that mute the potentially destabilizing roles of a few strong consumer–resource interactions.

The current empirical evidence indicates that communities may be dominated by such weak trophic interactions. If this is true, then it is also true that the removal, or addition, of any species (weak or strong) can lead to pronounced changes in community composition and structure. It follows that decreasing biodiversity will tend to increase the overall mean interaction strength, on average, and thus increase the probability that ecosystems undergo destabilizing dynamics and collapses. Just how much ecosystem deterioration is sufficient to precipitate a collapse is difficult to assess, but current experiments and theory agree that drastic community changes can accompany the removal or addition of even a single species. Furthermore, if Elton's observation is correct — that simplified communities are more vulnerable to invasion — than we should also expect an increase in frequency of successful invaders as well as an increase in their impact as our ecosystems become simplified. The lessons for conservation are obvious: (1) if we wish to preserve an ecosystem and its component species then we are best to proceed as if each species is sacred; and (2) species removals (that is, extinction) or species additions (that is, invasions) can, and eventually will, invoke major shifts in community structure and dynamics.

It is important to point out that the mechanistic ideas outlined here omit higher-scale ecosystem influences that are likely to be linked intricately to ecosystem stability and function^{70–73}. Some promising work is now beginning to show us how we can link models of nutrient and energy flow^{70–72} as well as uncover the potential influence of diversity and stability on large-scale biogeochemical processes⁷³. Investigations of this sort will be necessary to bridge important stabilizing processes that act across ecological scales. □

1. Ricciardi, A. & Rasmussen, J. B. Extinction rates of North American freshwater fauna. *Conserv. Biol.* **13**, 1220–1222 (2000).
2. Reid, W. V. Strategies for conserving biodiversity. *Environment* **39**, 16–43 (1997).
3. Levin, S. *Fragile Dominion: Complexity and the Commons* (Helix books, Reading, MA, 1999).
4. Lodge, D. Biological invasions: lessons for ecology. *Trends Ecol. Evol.* **8**, 133–137 (1993).
5. Cohen, A. & Carlton, J. T. Accelerating invasion rate in a highly invaded estuary. *Science* **279**, 555–558 (1998).
6. Odum, E. P. *Fundamentals of ecology* (Saunders, Philadelphia, 1953).
7. Elton, C. S. *Ecology of Invasions by Animals and Plants* (Chapman & Hall, London, 1958).
8. MacArthur, R. H. Fluctuations of animal populations and a measure of community stability. *Ecology* **36**, 533–536 (1955).
9. May, R. M. *Stability and complexity in model ecosystems* (Princeton Univ. Press, 1973).
10. Gardner, M. R. & Ashby, W. R. Connectance of large dynamic (cybernetic) systems: critical values for stability. *Nature* **228**, 784 (1970).
11. Pimm, S. L. & Lawton, J. H. On feeding on more than one trophic level. *Nature* **275**, 542–544 (1978).
12. Yodanis, P. The stability of real ecosystems. *Nature* **289**, 674–676 (1981).
13. Armstrong, R. A. & McGehee, R. Competitive exclusion. *Am. Nat.* **115**, 151–170 (1980).
14. DeAngelis, D. & Waterhouse, J. C. Equilibrium and nonequilibrium concepts in ecological models. *Ecol. Monogr.* **57**, 1–21 (1987).
15. Michalski, J. & Arditi, R. in *Advances in Environmental and Ecological Modelling* (ed. Weill, A.), 1–20 (Elsevier, Paris, 1999).
16. Huisman, J. & Weissing, F. J. Biodiversity of plankton by species oscillations and chaos. *Nature* **402**, 407–410 (1999).
17. McCann, K., Hastings, A. & Huxel, G. R. Weak trophic interactions and the balance of nature. *Nature* **395**, 794–798 (1998).
18. Law, R. & Morton, D. Permanence and the assembly of ecological communities. *Ecology* **77**, 762–775 (1996).
19. Hastings, A. & Higgins, K. Persistence of transients in spatially structured ecological models. *Science* **263**, 1133–1136 (1994).
20. Tilman, D. & Downing, J. A. Biodiversity and stability in grasslands. *Nature* **367**, 363–365 (1994).
21. Tilman, D., Wedin, D. & Knops, J. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* **379**, 718–720 (1996).
22. Tilman, D. Biodiversity: population versus ecosystem stability. *Ecology* **77**, 350–363 (1996).
23. Schapfer, F. & Schmid, B. Ecosystem effects of biodiversity: a classification of hypotheses and exploration of empirical results. *Ecol. Applic.* **9**, 893–912 (1999).
24. Doak, D. F. *et al.* The statistical inevitability of stability-diversity relationships in community ecology. *Am. Nat.* **151**, 264–276 (1998).
25. Tilman, D., Lehman, C. L. & Bristow, C. E. Diversity-stability relationships: statistical inevitability or ecological consequence. *Am. Nat.* **151**, 277–282 (1998).
26. Sankaran, M. & McNaughton, S. J. Determinants of biodiversity regulate compositional stability of communities. *Nature* **401**, 691–693 (1999).
27. Huston, M. A. Hidden treatments in ecological experiments: re-evaluating the ecosystem function of biodiversity. *Oecologia* **110**, 449–460 (1997).
28. Tilman, D. *et al.* The influence of functional diversity and composition on ecosystem processes. *Science* **277**, 1300–1302 (1997).
29. Hooper, D. U. & Vitousek, P. M. The effects of plant composition and diversity on ecosystem processes. *Science* **277**, 1302–1305 (1997).
30. Wardle, D. A. *et al.* Plant removals in perennial grassland: vegetation dynamics, decomposers, soil biodiversity, and ecosystem properties. *Ecol. Monogr.* **69**, 535–568 (1999).
31. van der Heijden, M. *et al.* Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396**, 69–72 (1998).
32. McNaughton, S. J. Ecology of a grazing ecosystem: the Serengeti. *Ecol. Monogr.* **55**, 259–294 (1985).
33. Lawton, J. H. Ecological experiments with model systems. *Science* **269**, 328–331 (1995).
34. McGrady-Steed, J., Harris, P. & Morin, P. J. Biodiversity regulates ecosystem predictability. *Nature* **390**, 162–165 (1997).
35. McGrady-Steed, J. & Morin, P. J. Biodiversity, density compensation, and the dynamics of populations and functional groups. *Ecology* **81**, 361–373 (2000).
36. Morin, P. J. & Lawler, S. P. Food web architecture and population dynamics: theory and empirical evidence. *Annu. Rev. Ecol. System.* **26**, 505–529 (1995).
37. Naeem, S. & Li, S. Biodiversity enhances ecosystem reliability. *Nature* **390**, 507–509 (1997).
38. Naeem, S. Species redundancy and ecosystem reliability. *Conserv. Biol.* **12**, 39–45 (1998).
39. Lawton, J. H. & Brown, V. K. in *Biodiversity and Ecosystem Function* (eds Schulze, E. D. & Mooney, H. A.), 255–270 (Springer, New York, 1993).
40. Yachi, S. & Loreau, M. Biodiversity and ecosystem functioning in a fluctuating environment: the insurance hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 1463–1468 (1999).
41. Chesson, P. & Huntley, N. The roles of harsh and fluctuating conditions in the dynamics of ecological communities. *Am. Nat.* **150**, 519–553 (1997).
42. Winemiller, K. O. Spatial and temporal variation in tropical fish trophic networks. *Ecol. Monogr.* **60**, 331–367 (1990).
43. Polis, G. A. Complex trophic interactions in deserts: an empirical critique of food web theory. *Am. Nat.* **138**, 123–155 (1991).
44. Polis, G. A. & Strong, D. Food web complexity and community dynamics. *Am. Nat.* **147**, 813–846 (1996).
45. Strong, D. Are trophic cascades all wet? Differentiation and donor-control in speciose ecosystems. *Ecology* **73**, 747–754 (1992).
46. Holt, R. D. in *Multitrophic interactions* (eds Begon, M., Gange, A. & Brown, V.) 333–350 (Chapman & Hall, London, 1996).
47. McCann, K. & Hastings, A. Re-evaluating the omnivory-stability relationship in food webs. *Proc. R. Soc. Lond. B* **264**, 1249–1254 (1997).
48. Huxel, G. R. & McCann, K. Food web stability: the influence of trophic flows across habitats. *Am. Nat.* **152**, 460–469 (1998).
49. Post, D. M., Connors, E. & Goldberg, D. S. Prey preference by a top predator and the stability of linked food chains. *Ecology* **81**, 8–14 (2000).
50. Yodanis, P. & Innes, S. Body-size and consumer-resource dynamics. *Am. Nat.* **139**, 1151–1175 (1992).
51. Chesson, J. The estimation and analysis of preference and its relationship to foraging models. *Ecology* **64**, 1297–1304 (1983).
52. Kokkoris, G. D., Troumbis, A. Y. & Lawton, J. H. Patterns of species interaction strength in assembled theoretical competition communities. *Ecol. Lett.* **2**, 70–74 (1999).
53. Paine, R. T. Food-web analysis through field measurements of per capita interaction strengths. *Nature* **355**, 73–75 (1992).
54. Fagan, W. F. & Hurd, L. E. Hatch density variation of a generalist arthropod predator: population consequences and community impact. *Ecology* **75**, 2022–2032 (1994).
55. Goldwasser, L. & Roughgarden, J. Construction and analysis of a large Caribbean food web. *Ecology* **74**, 1216–1223 (1993).
56. Raffaelli, D. G. & Hall, S. J. in *Food Webs: Integration of Patterns & Dynamics* (eds Polis, G. A. & Winemiller, K. O.) 185–191 (Chapman & Hall, New York, 1996).
57. Wootton, J. T. Estimates and tests of per capita interaction strength: diet abundance and impact of intertidally foraging birds. *Ecol. Monogr.* **67**, 45–64 (1997).
58. Paine, R. T. Ecological determinism in the competition for space. *Ecology* **65**, 1339–1348 (1984).
59. Berlow, E. Strong effects of weak interactions in ecological communities. *Nature* **398**, 330–334 (1999).
60. Williamson, M. & Fitter, A. The varying success of invaders. *Ecology* **77**, 1661–1666 (1996).
61. Vander Zanden, M. J., Casselman, J. M. & Rasmussen, J. B. Stable isotope evidence for the food web consequences of species invasions in lakes. *Nature* **401**, 464–467 (1999).
62. Fritts, T. H. & Rodda, G. H. The role of introduced species in the degradation of island ecosystems: a case history of Guam. *Annu. Rev. Ecol. System.* **29**, 113–140. (1998).
63. Reinthal, P. N. & Kling, G. W. in *Theory and Application in Fish Feeding Ecology* (eds Stouder, D. J., Fresh, K. L. & Feller, R.) 296–313 (Univ. South Carolina Press, 1994).
64. Fagan, W. F. Omnivory as a stabilizing feature of natural communities. *Am. Nat.* **150**, 554–567 (1997).
65. de Ruiter, P. C., Neutel, A. & Moore, J. C. Energetics, patterns of interaction strengths, and stability in real ecosystems. *Science* **269**, 1257–1260 (1995).
66. Holyoak, M. & Sachdev, S. Omnivory and the stability of simple food webs. *Oecologia* **117**, 413–419 (1999).
67. Flaherty, D. Ecosystem trophic complexity and densities of the Williamette mite, *Eotetranychus willamettei* ewing (Acarina: Tetranychidae). *Ecology* **50**, 911–916 (1969).
68. Morin, P. Productivity, intraguild predation, and population dynamics in experimental food webs. *Ecology* **80**, 752–760 (1999).
69. Luckinbill, L. S. Regulation, stability, and diversity in a model experimental microcosm. *Ecology* **60**, 1098–1102 (1979).
70. DeAngelis, D. *Dynamics of Nutrient Recycling and Food Webs* (Chapman & Hall, New York, 1992).
71. Andersen, T. *Pelagic Nutrient Cycles: Herbivores as Sources and Sinks* (Springer, New York, 1997).
72. Elser, J. J. & Urabe, J. The stoichiometry of consumer-driven nutrient recycling: theory, observations and consequences. *Ecology* **80**, 735–751 (1999).
73. Harding, S. P. Food web complexity enhances community stability and climate regulation in a geophysiological model. *Tellus* **51B**, 815–829 (1999).

Acknowledgements

This paper benefited from comments by D. Raffaelli. I also thank J. Rasmussen and P. Yodanis for conversations on this issue, and D. Kramer for providing a single comment that led me to a different viewpoint.

Consequences of changing biodiversity

F. Stuart Chapin III*, Erika S. Zavaleta†, Valerie T. Eviner§, Rosamond L. Naylor‡, Peter M. Vitousek†, Heather L. Reynolds||, David U. Hooper¶, Sandra Lavorel#, Osvaldo E. Sala☆, Sarah E. Hobbie**, Michelle C. Mack* & Sandra Díaz††

*Institute of Arctic Biology, University of Alaska, Fairbanks, Alaska 99775, USA (e-mail: fschapin@lter.uaf.edu)

†Department of Biological Sciences and ‡Institute for International Studies, Stanford University, Stanford, California 94305, USA

§Department of Integrative Biology, University of California, Berkeley, California 94720, USA

||Department of Biology, Kalamazoo College, Kalamazoo, Michigan 49006, USA

¶Department of Biology, Western Washington University, Bellingham, Washington 98225, USA

#Centre d'Ecologie Fonctionnelle et Evolutive, CNRS UPR 9056, 34293 Montpellier Cedex 05, France

☆Cátedra de Ecología and Instituto de Fisiología y Ecología Vinculadas a la Agricultura, Faculty of Agronomy, University of Buenos Aires, Ave San Martín 4453, Buenos Aires C1417DSE, Argentina

**Department of Ecology, Evolution, and Behavior, University of Minnesota, St Paul, Minnesota 55108, USA

††Instituto Multidisciplinario de Biología Vegetal, Universidad Nacional de Córdoba, FCEyN, Casilla de Correo 495, 5000 Córdoba, Argentina

Human alteration of the global environment has triggered the sixth major extinction event in the history of life and caused widespread changes in the global distribution of organisms. These changes in biodiversity alter ecosystem processes and change the resilience of ecosystems to environmental change. This has profound consequences for services that humans derive from ecosystems. The large ecological and societal consequences of changing biodiversity should be minimized to preserve options for future solutions to global environmental problems.

Humans have extensively altered the global environment, changing global biogeochemical cycles, transforming land and enhancing the mobility of biota. Fossil-fuel combustion and deforestation have increased the concentration of atmospheric carbon dioxide (CO₂) by 30% in the past three centuries (with more than half of this increase occurring in the past 40 years). We have more than doubled the concentration of methane and increased concentrations of other gases that contribute to climate warming. In the next century these greenhouse gases are likely to cause the most rapid climate change that the Earth has experienced since the end of the last glaciation 18,000 years ago and perhaps a much longer time. Industrial fixation of nitrogen for fertilizer and other human activities has more than doubled the rates of terrestrial fixation of gaseous nitrogen into biologically available forms. Run off of nutrients from agricultural and urban systems has increased several-fold in the developed river basins of the Earth, causing major ecological changes in estuaries and coastal zones. Humans have transformed 40–50% of the ice-free land surface, changing prairies, forests and wetlands into agricultural and urban systems. We dominate (directly or indirectly) about one-third of the net primary productivity on land and harvest fish that use 8% of ocean productivity. We use 54% of the available fresh water, with use projected to increase to 70% by 2050¹. Finally, the mobility of people has transported organisms across geographical barriers that long kept the biotic regions of the Earth separated, so that many of the ecologically important plant and animal species of many areas have been introduced in historic time^{2,3}.

Together these changes have altered the biological diversity of the Earth (Fig. 1). Many species have been eliminated from areas dominated by human influences. Even in

preserves, native species are often out-competed or consumed by organisms introduced from elsewhere. Extinction is a natural process, but it is occurring at an unnaturally rapid rate as a consequence of human activities. Already we have caused the extinction of 5–20% of the species in many groups of organisms (Fig. 2), and current rates of extinction are estimated to be 100–1,000 times greater than pre-human rates^{4,5}.

In the absence of major changes in policy and human behaviour, our effects on the environment will continue to alter biodiversity. Land-use change is projected to have the largest global impact on biodiversity by the year 2100, followed by climate change, nitrogen deposition, species introductions and changing concentrations of atmospheric CO₂ (ref. 6). Land-use change is expected to be of particular importance in the tropics, climatic change is likely to be important at high latitudes, and a multitude of interacting causes will affect other biomes (Fig. 3)⁶. What are the ecological and societal consequences of current and projected effects of human activity on biological diversity?

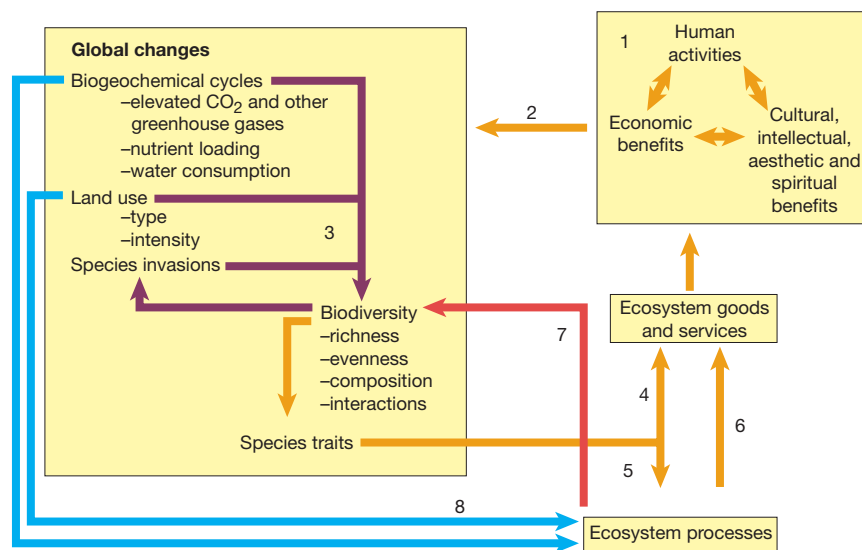
Ecosystem consequences of altered diversity

Diversity at all organizational levels, ranging from genetic diversity within populations to the diversity of ecosystems in landscapes, contributes to global biodiversity. Here we focus on species diversity, because the causes, patterns and consequences of changes in diversity at this level are relatively well documented. Species diversity has functional consequences because the number and kinds of species present determine the organismal traits that influence ecosystem processes. Species traits may mediate energy and material fluxes directly or may alter abiotic conditions (for example, limiting resources, disturbance and climate) that regulate process rates. The components of species diversity that determine this expression of traits include the number of species present (species richness), their relative abundances (species

Figure 1 The role of biodiversity in global change.

Human activities that are motivated by economic, cultural, intellectual, aesthetic and spiritual goals (1) are now causing environmental and ecological changes of global significance (2). By a variety of mechanisms, these global changes contribute to changing biodiversity, and changing biodiversity feeds back on susceptibility to species invasions (3, purple arrows; see text). Changes in biodiversity, through changes in species traits, can have direct consequences for ecosystem services and, as a result, human economic and social activities (4). In addition, changes in biodiversity can influence ecosystem

processes (5). Altered ecosystem processes can thereby influence ecosystem services that benefit humanity (6) and feedback to further alter biodiversity (7, red arrow). Global changes may also directly affect ecosystem processes (8, blue arrows). Depending on the circumstances, the direct effects of global change may be either stronger or weaker than effects mediated by changes in diversity. We argue that the costs of loss of biotic diversity, although traditionally considered to be 'outside the box' of human welfare, must be recognized in our accounting of the costs and benefits of human activities.



evenness), the particular species present (species composition), the interactions among species (non-additive effects), and the temporal and spatial variation in these properties. In addition to its effects on current functioning of ecosystems, species diversity influences the resilience and resistance of ecosystems to environmental change.

Species richness and evenness

Most theoretical and empirical work on the functional consequences of changing biodiversity has focused on the relationship between species richness and ecosystem functioning. Theoretical possibilities include positive linear and asymptotic relationships between richness and rates of ecosystem processes, or the lack of a simple statistical relationship⁷ (Box 1). In experiments, species richness correlates with rates of ecosystem processes most clearly at low numbers of species. We know much less about the impact of species richness in species-rich, natural ecosystems. Several studies using experimental species assemblages have shown that annual rates of primary productivity and nutrient retention increase with increasing plant species richness, but saturate at a rather low number of species^{8,9}. Arbuscular

mycorrhizal species richness also seems to enhance plant production in an asymptotic fashion, although phosphorus uptake was enhanced in a linear fashion from 1 to 14 species of fungi¹⁰. Microbial richness can lead to increased decomposition of organic matter¹¹. In contrast, no consistent statistical relationship has been observed between plant species richness of litter inputs and decomposition rate¹². Thus, in experimental communities (which typically focus on only one or two trophic levels), there seems to be no universal relationship between species richness and ecosystem functioning, perhaps because processes differ in their sensitivity to species richness compared with other components of diversity (such as evenness, composition or interactions). The absence of a simple relationship between species richness and ecosystem processes is likely when one or a few species have strong ecosystem effects.

Although the relationship of species richness to ecosystem functioning has attracted considerable theoretical and experimental attention because of the irreversibility of species extinction, human activities influence the relative abundances of species more frequently than the presence or absence of species. Changes in species evenness warrant increased attention, because they usually respond more rapidly to human activities than do changes in species richness and because they have important consequences to ecosystems long before a species is threatened by extinction.

Species composition

Particular species can have strong effects on ecosystem processes by directly mediating energy and material fluxes or by altering abiotic conditions that regulate the rates of these processes (Fig. 4)^{13,14}. Species' alteration of the availability of limiting resources, the disturbance regime, and the climate can have particularly strong effects on ecosystem processes. Such effects are most visible when introduced species alter previous patterns of ecosystem processes. For example, the introduction of the nitrogen-fixing tree *Myrica faya* to nitrogen-limited ecosystems in Hawaii led to a fivefold increase in nitrogen inputs to the ecosystem, which in turn changed most of the functional and structural properties of native forests¹⁵. Introduction of the deep-rooted salt cedar (*Tamarix* sp.) to the Mojave and Sonoran Deserts of North America increased the water and soil solutes

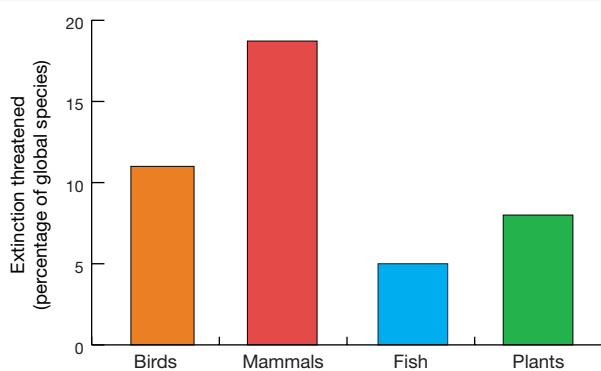


Figure 2 Proportion of the global number of species of birds, mammals, fish and plants that are currently threatened with extinction⁴.

Box 1

Species richness and ecosystem functioning

There has been substantial debate over both the form of the relationship between species richness and ecosystem processes and the mechanisms underlying these relationships⁸⁵. Theoretically, rates of ecosystem processes might increase linearly with species richness if all species contribute substantially and in unique ways to a given process — that is, have complementary niches. This relationship is likely to saturate as niche overlap, or ‘redundancy’, increases at higher levels of diversity⁸⁶. Several experiments indicate such an asymptotic relationship of ecosystem process rates with species richness. An asymptotic relationship between richness and process rates could, however, arise from a ‘sampling effect’ of increased probability of including a species with strong ecosystem effects, as species richness increases¹³. The sampling effect has at least two interpretations. It might be an important biological property of communities that influences process rates in natural ecosystems¹³, or it might be an artefact of species-richness experiments in which species are randomly assigned to treatments, rather than following community assembly rules that might occur in nature⁸⁷. Finally, ecosystem process rates may show no simple correlation with species richness. However, the lack of a simple statistical relationship between species richness and an ecosystem process may mask important functional relationships. This could occur, for example, if process rates depend strongly on the traits of certain species or if species interactions determine the species traits that are expressed (the ‘idiosyncratic hypothesis’)⁷. This mechanistic debate is important scientifically for understanding the functioning of ecosystems and effective management of their biotic resources. Regardless of the outcome of the debate, conserving biodiversity is essential because we rarely know *a priori* which species are critical to current functioning or provide resilience and resistance to environmental changes.

accessed by vegetation, enhanced productivity, and increased surface litter and salts. This inhibited the regeneration of many native species, leading to a general reduction in biodiversity¹⁶. The perennial tussock grass, *Agropyron cristatum*, which was widely introduced to the northern Great Plains of North America after the 1930s ‘dustbowl’, has substantially lower allocation to roots compared with native prairie grasses. Soil under *A. cristatum* has lower levels of available nitrogen and ~25% less total carbon than native prairie soil, so the introduction of this species resulted in an equivalent reduction of 480×10^{12} g carbon stored in soils¹⁷. Soil invertebrates, such as earthworms and termites, also alter turnover of organic matter and nutrient supply, thereby influencing the species composition of the aboveground flora and fauna¹⁸.

Species can also influence disturbance regime. For example, several species of nutritious but flammable grasses were introduced to the Hawaiian Islands to support cattle grazing. Some of these grasses spread into protected woodlands, where they caused a 300-fold increase in the extent of fire. Most of the woody plants, including some endangered species, are eliminated by fire, whereas grasses

rebound quickly¹⁹. Similar increases in the ecological role of fire resulting from grass invasions have been widely observed in the Americas, Australia and elsewhere in Oceania. The invasion of cheatgrass (*Bromus tectorum*) into western North America is one of the most extensive of these invasions. Cheatgrass has increased fire frequency by a factor of more than ten in the >40 million hectares ($1 \text{ ha} = 10^4 \text{ m}^2$) that it now dominates²⁰.

Species-induced changes in microclimate can be just as important as the direct impacts of environmental change. For example, in late-successional boreal forests, where soil temperatures have a strong influence on nutrient supply and productivity, the presence of moss, which reduces heat flux into the soil, contributes to the stability of permafrost (frozen soils) and the characteristically low rates of nutrient cycling²¹. As fire frequency increases in response to high-latitude warming, moss biomass declines, permafrost becomes less stable, the nutrient supply increases, and the species composition of forests is altered. Plant traits can also influence climate at larger scales. Simulations with general circulation models indicate that widespread replacement of deep-rooted tropical trees by shallow-

Figure 3 Scenarios of change in species diversity in selected biomes by the year 2100. The values are the projected change in diversity for each biome relative to the biome with greatest projected diversity change⁶. Biomes are: tropical forests (T), grasslands (G), Mediterranean (M), desert (D), north temperate forests (N), boreal forests (B) and arctic (A). Projected change in species diversity is calculated assuming three alternative scenarios of interactions among the causes of diversity change. Scenario 1 assumes no interaction among causes of diversity change, so that the total change in diversity is the sum of the changes caused by each driver of diversity change. Scenario 2 assumes that only the factor with the greatest impact on diversity influences diversity change. Scenario 3 assumes that factors causing change in biodiversity interact multiplicatively to determine diversity change. For scenarios 1 and 2, we show the relative importance of the major causes of projected change in diversity. These causes are climatic change, change in land use, introduction of exotic species, and changes in atmospheric CO₂ and/or nitrogen deposition (labelled ‘other’). The graph shows that all biomes are projected to experience substantial change in species diversity by 2100, that the most important causes of diversity change differ among biomes, and that the patterns of diversity change depend on assumptions about the nature of interactions among the causes of diversity change. Projected biodiversity change is most similar among biomes if causes of diversity change do not interact (scenario 1) and differ most strongly among biomes if the causes of biodiversity change interact multiplicatively (scenario 3).

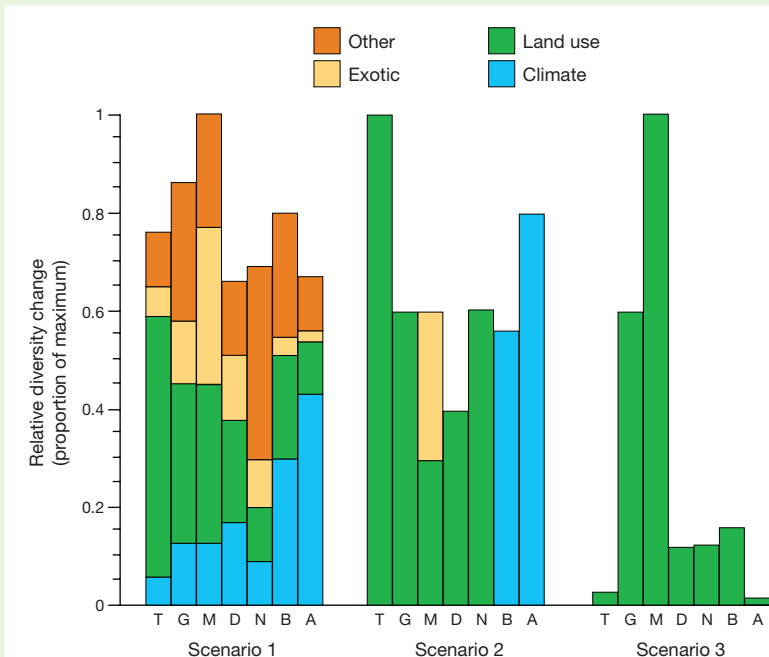
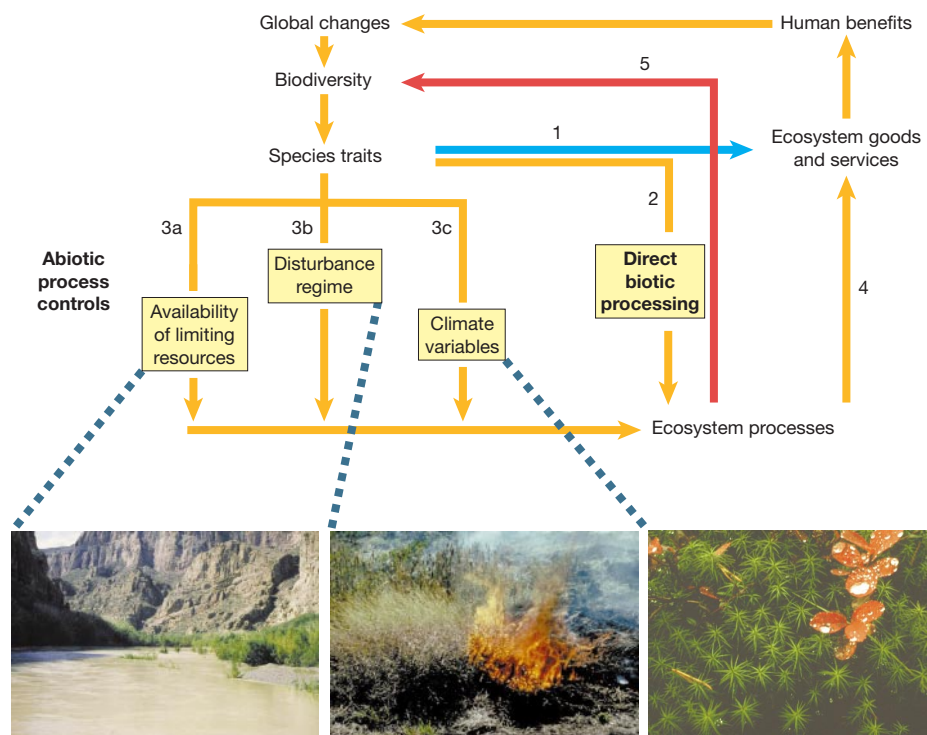


Figure 4 Mechanisms by which species traits affect ecosystem processes. Changes in biodiversity alter the functional traits of species in an ecosystem in ways that directly influence ecosystem goods and services (1) either positively (for example, increased agricultural or forestry production) or negatively (for example, loss of harvestable species or species with strong aesthetic/cultural value). Changes in species traits affect ecosystem processes directly through changes in biotic controls (2) and indirectly through changes in abiotic controls, such as availability of limiting resources (3a), disturbance regime (3b), or micro- or macroclimate variables (3c). Illustrations of these effects include: reduction in river flow due to invasion of deep-rooted desert trees (3a; photo by E. Zavaleta); increased fire frequency resulting from grass invasion that destroys native trees and shrubs in Hawaii (3b, photo by C. D'Antonio); and insulation of soils by mosses in arctic tundra, contributing to conditions that allow for permafrost (3c; photo by D. Hooper). Altered processes can then influence the availability of ecosystem goods and services directly (4) or indirectly by further altering biodiversity (5), resulting in loss of useful species or increases in noxious species.



rooted pasture grasses would reduce evapotranspiration and lead to a warmer, drier climate²². At high latitudes, the replacement of snow-covered tundra by a dark conifer canopy will probably increase energy absorption sufficiently to act as a powerful positive feedback to regional warming²³.

Species interactions

Most ecosystem processes are non-additive functions of the traits of two or more species, because interactions among species, rather than simple presence or absence of species, determine ecosystem characteristics (Fig. 5). Species interactions, including mutualism, trophic interactions (predation, parasitism and herbivory), and competition may affect ecosystem processes directly by modifying pathways of energy and material flow²⁴ or indirectly by modifying the abundances or traits of species with strong ecosystem effects²⁵.

Mutualistic species interactions contribute directly to many essential ecosystem processes. For example, nitrogen inputs to terrestrial ecosystems are mediated primarily by mutualistic associations between plants and nitrogen-fixing microorganisms. Mycorrhizal associations between plant roots and fungi greatly aid plant nutrient uptake from soil, increase primary production and speed succession²⁶. Highly integrated communities (consortia) of soil microorganisms, in which each species contributes a distinct set of enzymes, speeds the decomposition of organic matter²⁷. Many of these interactions have a high degree of specificity, which increases the probability that loss of a given species will have cascading effects on the rest of the system.

Trophic interactions can have large effects on ecosystem processes either by directly modifying fluxes of energy and materials, or by influencing the abundances of species that control those fluxes. When top predators are removed, prey populations sometimes explode and deplete their food resources, leading to a cascade of ecological effects. For example, removal of sea otters by Russian fur traders allowed a population explosion of sea urchins that

overgrazed kelp²⁸ (Fig. 6a). Recent over-fishing in the North Pacific may have triggered similar outbreaks of sea urchin, as killer whales moved closer to shore and switched to sea otters as an alternate prey²⁹. In the absence of dense populations of sea urchins, kelp provides the physical structure for diverse subtidal communities and attenuates waves that otherwise augment coastal erosion and storm damage³⁰. Removing bass from lakes that were fertilized with phosphorus caused an increase in minnows, which depleted the biomass of phytoplankton grazers and caused algal blooms³¹ (Fig. 6b). The algal blooms turned the lake from a net source to a net sink of CO₂. Thus, biotic change and altered nutrient cycles can interact to influence whole-system carbon balance. The zebra mussel (*Dreissena polymorpha*) is a bottom-dwelling invasive species that, through its filter feeding, markedly reduces phytoplankton while increasing water clarity and phosphorus availability³². Introduction of this species shifts the controlling interactions of the food web from the water column to the sediments. Trophic interactions are also important in terrestrial ecosystems. At the micro scale, predation on bacteria by protozoan grazers speeds nitrogen cycling near plant roots, enhancing nitrogen availability to plants³³. At the regional scale, an improvement in hunting technology at the end of the Pleistocene may have contributed to the loss of the Pleistocene megafauna and the widespread change from steppe grassland to tundra that occurred in Siberia 10,000–18,000 years ago³⁴. The resulting increase in mosses insulated the soil and led to cooler soils, less decomposition and greater sequestration of carbon in peat. Today, human harvest of animals continues to have a pronounced effect of the functioning of ecosystems.

Competition, mutualisms and trophic interactions frequently lead to secondary interactions among other species, often with strong ecosystem effects (Fig. 5). For example, soil microbial composition can modify the outcome of competition among plant species³⁵, and plants modify the microbial community of their

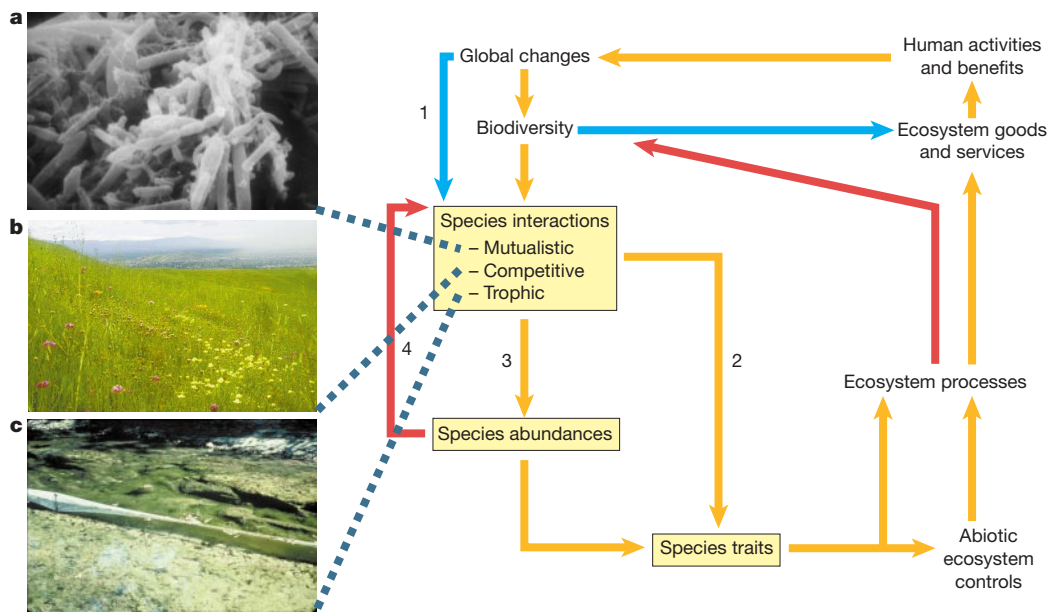


Figure 5 Mechanisms by which species interactions affect ecosystem processes. Global environmental change affects species interactions (mutualism, competition and trophic interactions) both directly (1) and through its effects on altered biodiversity. Species interactions may directly affect key traits (for example, the inhibition of microbial nitrogen fixation by plant secondary metabolites) in ecosystem processes (2) or may alter the abundances of species with key traits (3). Examples of these species interactions include (a) mutualistic consortia of microorganisms, each of which produces only some of the enzymes required to break down organic matter (photo by M. Klug), (b) altered abundances of native California forbs due to competition from introduced European grasses (photo by H. Reynolds), and (c) alteration of algal biomass due to presence or absence of grazing minnows⁸⁴ (photo by M. Power). Changes in species interactions and the resulting changes in community composition (3) may feedback to cause a cascade of further effects on species interactions (4).

neighbours, which, in turn, affects nitrogen supply and plant growth³⁶. Stream predatory invertebrates alter the behaviour of their prey, making them more vulnerable to fish predation, which leads to an increase in the weight gain of fish³⁷. In the terrestrial realm, grazers can reduce grass cover to the point that avian predators keep vole populations at low densities, allowing the persistence of *Erodium botrys*, a preferred food of voles³⁸. The presence of *E. botrys* increases leaching³⁹ and increases soil moisture⁴⁰, which often limits production and nutrient cycling in dry grasslands. These examples clearly indicate that all types of organisms — plants, animals and microorganisms — must be considered in understanding the effects of biodiversity on ecosystem functioning. Although each of these examples is unique to a particular ecosystem, the ubiquitous nature of species interactions with strong ecosystem effects makes these interactions a general feature of ecosystem functioning. In many cases, changes in these interactions alter the traits that are expressed by species and therefore the effects of species on ecosystem processes. Consequently, simply knowing that a species is present or absent is insufficient to predict its impact on ecosystems.

Many global changes alter the nature or timing of species interactions⁴¹. For example, the timing of plant flowering and the emergence of pollinating insects differ in their responses to warming, with potentially large effects on ecosystems and communities⁴². Plant–herbivore interactions in diverse communities are less likely to be disrupted by elevated CO₂ (ref. 43) than in simple systems involving one specialist herbivore and its host plant⁴⁴.

Resistance and resilience to change

The diversity–stability hypothesis suggests that diversity provides a general insurance policy that minimizes the chance of large ecosystem changes in response to global environmental change⁴⁵. Microbial microcosm experiments show less variability in ecosystem processes in communities with greater species richness⁴⁶, perhaps because every species has a slightly different response to its physical and biotic environment. The larger the number of functionally similar species

in a community, the greater is the probability that at least some of these species will survive stochastic or directional changes in environment and maintain the current properties of the ecosystem⁴⁷. This stability of processes has societal relevance. Many traditional farmers plant diverse crops, not to maximize productivity in a given year, but to decrease the chances of crop failure in a bad year⁴⁸. Even the loss of rare species may jeopardize the resilience of ecosystems. For example, in rangeland ecosystems, rare species that are functionally similar to abundant ones become more common when grazing reduces their abundant counterparts. This compensation in response to release from competition minimizes the changes in ecosystem properties⁴⁹.

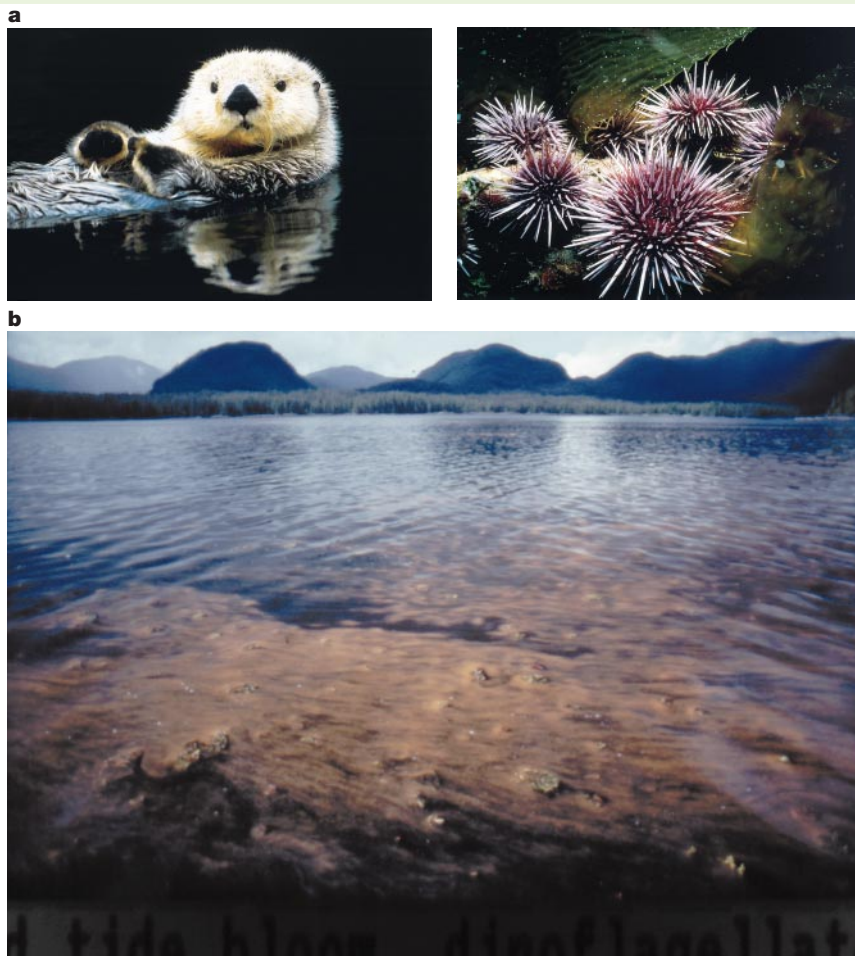
Species diversity also reduces the probability of outbreaks by ‘pest’ species by diluting the availability of their hosts. This decreases host-specific diseases⁵⁰, plant-feeding nematodes⁵¹ and consumption of preferred plant species⁵². In soils, microbial diversity decreases fungal diseases owing to competition and interference among microbes⁵³.

Resistance to invasions

Biodiversity can influence the ability of exotic species to invade communities through either the influence of traits of resident species or some cumulative effect of species richness. Early theoretical models and observations of invasions on islands indicated that species-poor communities would be more vulnerable to invasions because they offered more empty niches⁵⁴. However, studies of intact ecosystems find both negative⁵⁵ and positive⁵⁶ correlations between species richness and invasions. This occurs in part because the underlying factors that generate differences in diversity (for example, propagule supply, disturbance regime and soil fertility) cannot be controlled and may themselves be responsible for differences in invasibility⁵⁶. The diversity effects on invasibility are scale-dependent in some cases. For example, at the plot scale, where competitive interactions might exert their effect, increased plant diversity correlated with lower vulnerability to invasion in Central Plains grasslands of the United States. Across landscape scales, however, ecological factors that promote

Figure 6 Trophic interactions can affect ecosystem processes by influencing species' abundances.

a, Removal of sea otters by Russian fur traders caused an explosion in the population of sea urchins that overgrazed kelp. (Photographs courtesy of M. Sewell/Still Pictures and J. Rotman/BBC Natural History Unit.) **b**, Similarly, changes in the species balance and the abundance of fish can deplete phytoplankton grazers and cause algal blooms. (Photograph courtesy of J. Foott/BBC Natural History Unit.)



native plant diversity (for example, soil type and disturbance regime) also promote species invasions⁵⁷.

Experimental studies with plants⁵⁸ or soil microorganisms⁵⁹ often show that vulnerability to invasion is governed more strongly by the traits of resident and invading species than by species richness *per se*. Both competition and trophic interactions contribute to these effects of community composition on invasibility. For example, in its native range, the Argentine ant (*Linepithaema humile*) is attacked by species-specific parasitoids that modify its behaviour and reduce its ability to dominate food resources and competitively exclude other ant species⁶⁰. These parasitoids are absent from the introduced range of Argentine ants, which may explain their success at eliminating native ant communities in North America⁶¹. Observational and experimental studies together indicate that the effect of species diversity on vulnerability to invasion depends on the components of diversity involved (richness, evenness, composition and species interactions) and their interactions with other ecological factors such as disturbance regime, resource supply and rate of propagule arrival. Humans significantly affect all of these factors (Figs 1, 4), thereby dramatically increasing the incidence of invasions worldwide.

Societal consequences of altered diversity

Biodiversity and its links to ecosystem properties have cultural, intellectual, aesthetic and spiritual values that are important to society. In addition, changes in biodiversity that alter ecosystem functioning have economic impacts through the provision of ecosystem goods and services to society (Fig. 1 and Box 2). Changes in diversity can directly reduce sources of food, fuel, structural materials, medicinals or genetic resources. These changes can also alter the abundance of other species that control ecosystem processes, leading to further

changes in community composition and vulnerability to invasion. Introduction of exotic species or changes in community composition can affect ecosystem goods or services either by directly reducing abundances of useful species (by predation or competition), or by altering controls on critical ecosystem processes (Fig. 4).

These impacts can be wide-ranging and costly. For example, the introduction of deep-rooted species in arid regions reduces supplies and increases costs of water for human use. Marginal water losses to the invasive star thistle, *Centaurea solstitialis*, in the Sacramento River valley, California, have been valued at US\$16–56 million per year (J. D. Gerlach, unpublished results) (Fig. 7). In South Africa's Cape region, the presence of rapidly transpiring exotic pines raises the unit cost of water procurement by nearly 30% (ref. 62). Increased evapotranspiration due to the invasion of *Tamarix* in the United States costs an estimated \$65–180 million per year in reduced municipal and agricultural water supplies⁶³. In addition to raising water costs, the presence of sediment-trapping *Tamarix* stands has narrowed river channels and obstructed over-bank flows throughout the western United States, increasing flood damages by as much as \$50 million annually⁶³.

Those species changes that have greatest ecological impact frequently incur high societal costs. Changes in traits maintaining regional climate²² constitute an ecosystem service whose value in tropical forests has been estimated at \$220 ha⁻¹ yr⁻¹ (ref. 64). The loss or addition of species that alter disturbance regimes can also be costly. The increased fire frequency resulting from the cheatgrass invasion in the western United States has reduced rangeland values and air quality and led to increased expenditures on fire suppression⁶⁵. The disruption of key species interactions can also have large societal and ecological consequences. Large populations of passenger pigeons (*Ectopistes migratorius*) in the northeastern United States

may once have controlled Lyme tick-bearing mice by out-competing them for food⁶⁶. The loss of the passenger pigeon to nineteenth-century over-hunting may, therefore, have contributed to the rise of Lyme disease in humans in the twentieth century. The economic impacts of invasions of novel species are particularly well documented. The introduction and spread of single pests such as the golden apple snail (*Pomacea canaliculata*) and the European corn borer (*Ostrinia nubilalis*) have had major impacts on food production and farm incomes^{67,68}. Estimates of the overall cost of invasions by exotic species in the United States range widely from \$1.1 to \$137 billion annually^{69,70}. In Australia, plant invasions alone entail an annual cost of US\$2.1 billion⁷¹.

The provision of tangible ecosystem goods and services by natural systems depends not only on species' presence or absence but also on their abundance. Large populations of the white-footed mouse (*Peromyscus leucopus*) in the northeastern United States control outbreaks of gypsy moth (*Lymantria dispar*) but spread Lyme disease, whereas small populations of the mouse decrease the incidence of Lyme disease but allow gypsy moth defoliation⁷². An analysis of the costs of changes in biodiversity thus involves more than just analysis of extinctions and invasions. The loss of a species to extinction is of special societal concern, however, because it is irreversible. Future opportunities to learn and derive newly recognized benefits from an extinct species are lost forever. Preventing such a loss preserves an 'option value' for society — the value of attaining more knowledge about species and their contribution to human well being in order to make informed decisions in the future^{73,74}. For example, significant value (\$230–330 million) has been attributed to genetic information gained from preventing land conversion in Jalisco, Mexico, in an area containing a wild grass, teosinte (*Euchlaena mexicana*), that can be used to develop viral-resistant strains of perennial corn⁷³. If this land had been converted to agriculture or human settlements, the societal benefits of development would have come at the expense of an irreversible loss in genetic material that could be used for breeding viral resistance in one of the most widely consumed cereal crops in the world. The perceived costs of diversity loss in this situation might have been small — especially relative to the development benefits — whereas the actual (unrecognized) costs of losing genetic diversity would have been significant (Fig. 8). Decisions to preserve land to gain further information about the societal value of species diversity or ecosystem function typically involve a large degree of uncertainty, which often leads to myopic decisions regarding land use.

Figure 7 Water losses to the invasive, deep-rooted star thistle, *C. solstitialis*, provides an example of the financial impacts of introducing exotic species on ecosystem composition. (Photograph courtesy of P. Collins/A-Z Botanical Collection.)



Box 2

Ecosystem services

Ecosystem services are defined as the processes and conditions of natural ecosystems that support human activity and sustain human life. Such services include the maintenance of soil fertility, climate regulation and natural pest control, and provide flows of ecosystem goods such as food, timber and fresh water. They also provide intangible benefits such as aesthetic and cultural values⁸⁸.

Ecosystem services are generated by the biodiversity present in natural ecosystems. Ecologists and economists have begun to quantify the impacts of changes in biodiversity on the delivery of ecosystem services and to attach monetary value to these changes. Techniques used to attach value to biodiversity change range from direct valuation based on market prices to estimates of what individuals are willing to pay to protect endangered wildlife⁸⁹.

Although there are estimates of the global values of ecosystem services⁶⁴, valuation of the marginal losses that accompany specific biodiversity changes are most relevant to policy decisions.

Predicting the value of such losses involves uncertainty, because ecological and societal systems interact in nonlinear ways and because human preferences change through time. Assumptions today about future values may underestimate the values placed on natural systems by future generations⁸⁹. Therefore, minimizing loss of biodiversity offers a conservative strategy for maintaining this value.

Global environmental changes have the potential to exacerbate the ecological and societal impacts of changes in biodiversity⁶. In many regions, land conversion forces declining populations towards the edges of their species range, where they become increasingly vulnerable to collapse if exposed to further human impact⁷⁵. Warming allows the poleward spread of exotics and pathogens, such as dengue- and malaria-transmitting mosquitoes (*Aedes* and *Anopheles* sp.)⁷⁶ and pests of key food crops, such as corn-boring insects⁶⁸. Warming can also exacerbate the impacts of water-consuming invasive plant species in water-scarce areas by increasing regional water losses. The *Tamarix*-invaded Colorado River in the United States currently has a mean annual flow that is 10% less than regional water allocations for human use⁷⁷. Warming by 4°C would reduce the flow of the Colorado River by more than 20%, further increasing the marginal costs of water losses to *Tamarix*⁷⁸. Similar impacts of global change in regions such as Sahelian Africa, which have less water and less well developed distribution mechanisms, might directly affect human survival. In many cases, accelerated biodiversity loss is already jeopardizing the livelihoods of traditional peoples⁷⁹.

The combination of irreversible species losses and positive feedbacks between biodiversity changes and ecosystem processes are likely to cause nonlinear cost increases to society in the future, particularly when thresholds of ecosystem resilience are exceeded⁸⁰. For example, *Imperata cylindrica*, an aggressive indigenous grass, colonizes forest lands of Asia that are cleared for slash-and-burn agriculture, forming a monoculture grassland with no vascular plant diversity and many fewer mammalian species than the native forest. The total area of *Imperata* in Asia is currently about 35 million ha (4% of land area)⁸¹. Once in place, *Imperata* is difficult and costly to remove and enhances fire, which promotes the spread of the grass. The annual cost of reversing this conversion in Indonesia, where 4% of the nation's area (8.6 million ha) is now in *Imperata* grasslands, would be over \$400 million if herbicides are used, and \$1.2 billion if labour is used to remove the grass manually. Farmers typically burn the fields because herbicides and labour are too expensive. Burning these grasslands, however, increases losses of soil nitrogen and carbon, which erode agricultural productivity, and enhances regeneration of *Imperata*. This positive feedback with nonlinear changes in land cover will probably continue in the future as lands are deforested

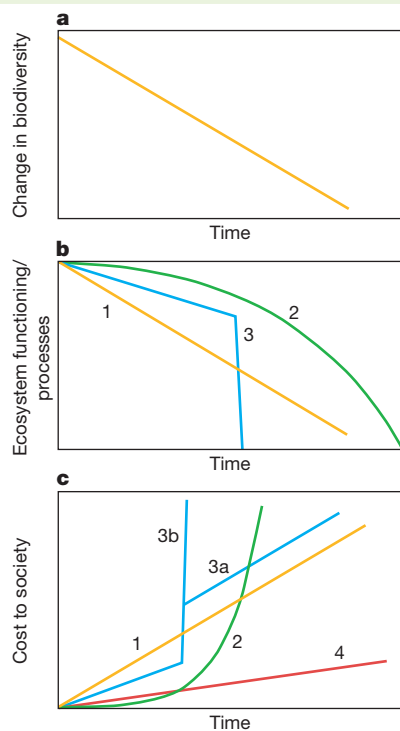


Figure 8 Ecosystem and societal consequences of changes in biodiversity. **a**, A linear change in biodiversity through time. **b**, This change might (1) induce a linear response in ecosystem processes, (2) have increasingly large impacts on ecosystem functioning, yielding exponential ecosystem change through time, or (3) exhibit abrupt thresholds owing to the loss of a keystone species, the loss of the last member of a key functional group, or the addition of a new species trait. **c**, Even if ecosystem response to diversity changes is linear, associated societal costs through time may respond nonlinearly. Departures from a linear increase (1) in societal costs over time might include larger cost increases (2) associated with each additional unit of change in ecosystem processes, yielding an exponential cost curve through time. Reductions of resource supply below threshold levels may induce step increases in societal costs (3a), such as reductions in water supply below the point where all consumers have access to enough for desired uses. If changes in resource supply or ecosystem processes exceed thresholds for supporting large segments of society, stepwise cost increases may be unmeasurable or essentially infinite (3b). The perceived ecological changes and societal costs of diversity change may be small (4). Actual, unrecognized costs may be far higher (lines 1, 2 and 3) and discovered only later as lost option values. Conservation of biodiversity can help avoid such negative ecological and economic 'surprises'.

for timber and agricultural purposes, causing further declines in regional biodiversity.

Uncertainty related to positive feedbacks and nonlinear changes in land cover and biodiversity make social adaptation to change more difficult and costly (Fig. 8). It may be more important from an economic perspective to understand the nature and timing of rapid or nonlinear changes in societal costs caused by loss of biodiversity and associated ecosystem services than it is to predict average consequences of current trends of species decline. By analogy, economic models of ecological 'surprises' in response to climatic change show that the information about the nonlinearities in damage from warming is worth up to six times more than information about current trends in damage levels⁸². In the *Imperata* example, the costs of replacing the original ecosystem goods and services from the forest — including timber products, fire stability and soil nutrients — rise sharply as *Imperata* spreads. If these nonlinearities in the ecological and economic effects of this conversion had been anticipated, policies could have been implemented to encourage agroforestry instead of rice production or to reduce migration and settlement in

the most vulnerable areas⁸³.

In sum, these examples indicate a tight coupling between altered species diversity, ecosystem function and societal costs. A pressing task for ecologists, land managers and environmental policy makers is to determine where and when such tight couplings exist. Policies to safeguard ecosystem services must be able to respond dynamically to new knowledge, the rapidly changing global environment, and evolving societal needs. Nonlinearity, uncertainty and irreversibility call for a more aggressive approach to mitigating changes in biodiversity than is now being pursued so that future options are not foreclosed.

Conclusion

We are in the midst of one of the largest experiments in the history of the Earth. Human effects on climate, biogeochemical cycles, land use and mobility of organisms have changed the local and global diversity of the planet, with important ecosystem and societal consequences (Fig. 1). The most important causes of altered biodiversity are factors that can be regulated by changes in policy: emissions of greenhouse gases, land-use change and species introductions. In the past, the international community has moved to reduce detrimental human impacts with unambiguous societal consequences. For example, the Montreal Protocol prohibited release of chlorofluorocarbons in response to evidence that these chemicals caused loss of ozone and increased levels of cancer-producing UV-B radiation. Strong evidence for changes in biodiversity and its ecosystem and societal consequences calls for similar international actions. We urge the following blueprint for action.

- The scientific community should intensify its efforts to identify the causes of nonlinearities and thresholds in the response of ecosystem and social processes to changes in biodiversity.
- The scientific community and informed citizens should become engaged in conveying to the public, policy-makers and land managers the enormity and irreversibility of current rapid changes in biodiversity. Despite convincing scientific evidence, there is a general lack of public awareness that change in biodiversity is a global change with important ecological and societal impacts and that these changes are not amenable to mitigation after they have occurred.
- Managers should consider the ecological and social consequences of biodiversity change at all stages in land-use planning. For example, environmental impact assessments should consider both the current costs of ecosystem services that will be lost and the risk of nonlinear future change. Managed landscapes can support a large proportion of regional biodiversity with proper planning, management and adaptive responses.
- Scientists and other citizens should collaborate with governmental organizations, from local to national levels, in developing and implementing policies and regulations that reduce environmental deterioration and changes in biodiversity. For example, more stringent restrictions on the import of biotic materials could curb the rate of biotic invasions, and improved land and watershed management could reduce their rates of spread.
- A new international body that would be comparable to the Intergovernmental Panel on Climate Change (IPCC) should assess changes in biodiversity and their consequences as an integral component of the assessment of the societal impacts of global change.
- International bodies should establish and implement agreements such as the Convention on Biological Diversity that institute mechanisms for reducing activities that drive the changes in biodiversity. These activities include fossil-fuel emissions, land-use change and biotic introductions. □

1. Postel, S. L., Daily, G. C. & Ehrlich, P. R. Human appropriation of renewable fresh water. *Science* **271**, 785–788 (1996).
2. Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. Human domination of Earth's ecosystems. *Science* **277**, 494–499 (1997).
3. Kattenberg, A. et al. in *Climate Change 1995. The Science of Climate Change* (ed. Houghton, J. T.) 285–357 (Cambridge Univ. Press, Cambridge, 1996).

4. Pimm, S. L., Russell, G. J., Gittleman, J. L. & Brooks, T. M. The future of biodiversity. *Science* **269**, 347–350 (1995).
5. Lawton, J. H. & May, R. M. *Extinction Rates* (Oxford Univ. Press, Oxford, 1995).
6. Sala, O. E. *et al.* Global biodiversity scenarios for the year 2100. *Science* **287**, 1770–1776 (2000).
7. Lawton, J. H. What do species do in ecosystems? *Oikos* **71**, 367–374 (1994).
8. Tilman, D., Wedin, D. & Knops, J. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature* **379**, 718–720 (1996).
9. Hector, A. *et al.* Plant diversity and productivity experiments in European grasslands. *Science* **286**, 1123–1127 (1999).
10. van der Heijden, M. G. A. *et al.* Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396**, 69–72 (1998).
11. Saloniemi, P. O. Metabolic capabilities of forest soil microbial populations with reduced species diversity. *Soil Biol. Biochem.* **13**, 1–10 (1981).
12. Wardle, D. A., Bonner, K. I. & Nicholson, K. S. Biodiversity and plant litter: experimental evidence which does not support the view that enhanced species richness improves ecosystem function. *Oikos* **79**, 247–258 (1997).
13. Hooper, D. U. & Vitousek, P. M. The effects of plant composition and diversity on ecosystem processes. *Science* **277**, 1302–1305 (1997).
14. Tilman, D. *et al.* The influence of functional diversity and composition on ecosystem processes. *Science* **277**, 1300–1302 (1997).
15. Vitousek, P. M., Walker, L. R., Whiteaker, L. D., Mueller-Dombois, D. & Matson, P. A. Biological invasion by *Myrica faya* alters ecosystem development in Hawaii. *Science* **238**, 802–804 (1987).
16. Berry, W. L. Characteristics of salts secreted by *Tamarix aphylla*. *Am. J. Bot.* **57**, 1226–1230 (1970).
17. Christian, J. M. & Wilson, S. D. Long-term impacts of an introduced grass in the northern Great Plains. *Ecology* **80**, 2397–2407 (1999).
18. Lavelle, P., Bignell, D. & Lepage, M. Soil function in a changing world: the role of invertebrate ecosystem engineers. *Eur. J. Soil Biol.* **33**, 159–193 (1997).
19. D'Antonio, C. M. & Vitousek, P. M. Biological invasions by exotic grasses, the grass-fire cycle, and global change. *Annu. Rev. Ecol. Syst.* **23**, 63–87 (1992).
20. Whisenant, S. *Changing Fire Frequencies on Idaho's Snake River Plains: Ecological Management Implications* 4–10 (US Forest Service General Technical Report INT-276, Washington, 1990).
21. Van Cleave, K., Chapin, F. S. III, Dryness, C. T. & Viereck, L. A. Element cycling in taiga forest: state-factor control. *BioScience* **41**, 78–88 (1991).
22. Shukla, J., Nobre, C. & Sellers, P. Amazon deforestation and climate change. *Science* **247**, 1322–1325 (1990).
23. Foley, J. A., Kutzbach, J. E., Coe, M. T. & Levis, S. Feedbacks between climate and boreal forests during the Holocene epoch. *Nature* **371**, 52–54 (1994).
24. de Ruiter, P. C., Neutel, A. & Moore, J. C. Energetics, patterns of interaction strengths, and stability in real ecosystems. *Science* **269**, 1257–1260 (1995).
25. Power, M. E. *et al.* Challenges in the quest for keystones. *BioScience* **46**, 609 (1996).
26. Read, D. J. Mycorrhizas in ecosystems. *Experientia* **47**, 376–391 (1991).
27. Paerl, H. W. & Pinckney, J. L. A mini-review of microbial consortia: their roles in aquatic production and biogeochemical cycling. *Microbial Ecol.* **31**, 225–247 (1996).
28. Estes, J. A. & Palmisano, J. F. Sea otters: their role in structuring nearshore communities. *Science* **185**, 1058–1060 (1974).
29. Estes, J. A., Tinker, M. T., Williams, T. M. & Doak, D. F. Killer whale predation on sea otters linking oceanic and nearshore ecosystems. *Science* **282**, 473–476 (1998).
30. Mork, M. The effect of kelp in wave damping. *Sarsia* **80**, 323–327 (1996).
31. Schindler, D. E., Carpenter, S. R., Cole, J. J., Kitchell, J. F. & Pace, M. L. Influence of food web structure on carbon exchange between lakes and the atmosphere. *Science* **277**, 248–251 (1997).
32. Caraco, N. F. *et al.* Zebra mussel invasion in a large, turbid river: phytoplankton response to increased grazing. *Ecology* **78**, 588–602 (1997).
33. Clarholm, M. Interactions of bacteria, protozoa and plants leading to mineralization of soil nitrogen. *Soil Biol. Biochem.* **17**, 181–187 (1985).
34. Zimov, S. A. *et al.* Steppe-tundra transition: an herbivore-driven biome shift at the end of the Pleistocene. *Am. Nat.* **146**, 765–794 (1995).
35. Chanway, C. P., Turkington, R. & Holl, F. B. Ecological implications of specificity between plants and rhizosphere micro-organisms. *Adv. Ecol. Res.* **21**, 121–169 (1991).
36. Lawley, R. A., Newman, E. I. & Campbell, R. Abundance of endomycorrhizas and root-surface microorganisms on three grasses grown separately and in mixtures. *Soil Biol. Biochem.* **14**, 237–240 (1982).
37. Soluck, D. A. & Richardson, J. S. The role of stoneflies in enhancing growth of trout: a test of the importance of predator–predator facilitation within a stream community. *Oikos* **80**, 214–219 (1997).
38. Rice, K. J. Interaction of disturbance patch size and herbivory in *Erodium* colonization. *Ecology* **68**, 1113–1115 (1987).
39. Shock, C. C., Jones, M. B., Williams, W. A. & Center, D. M. Competition of S and N by associations of three annual range species in lysimeters. *Plant Soil* **81**, 311–321 (1984).
40. Gordon, D. R. & Rice, K. J. Partitioning of space and water between two California annual grassland species. *Am. J. Bot.* **79**, 967–976 (1992).
41. Jifon, J. L., Friend, A. L. & Berrang, P. C. Species mixture and soil-resource availability affect the root growth response of tree seedlings to elevated atmospheric CO₂. *Can. J. For. Res.* **25**, 824–832 (1995).
42. Harrington, R., Woiwod, I. & Sparks, T. Climate change and trophic interactions. *Trends Ecol. Evol.* **14**, 146–150 (1999).
43. Diaz, S., Fraser, L. H., Grime, J. P. & Falczuk, V. The impact of elevated CO₂ on plant–herbivore interactions: experimental evidence of moderating effects at the community level. *Oecologia* **117**, 177–186 (1998).
44. Lindroth, R. L. in *Carbon Dioxide and Terrestrial Ecosystems* (eds Koch, G. W. & Mooney, H. A.) 105–120 (Academic Press, San Diego, 1996).
45. McNaughton, S. J. Diversity and stability of ecological communities: a comment on the role of empiricism in ecology. *Am. Nat.* **111**, 515–525 (1977).
46. Naeem, S. & Li, S. Biodiversity enhances ecosystem reliability. *Nature* **390**, 507–509 (1997).
47. Chapin, F. S. III & Shaver, G. R. Individualistic growth response of tundra plant species to environmental manipulations in the field. *Ecology* **66**, 564–576 (1985).
48. Altieri, M. A. in *Agroecology* (eds Carrol, C. R., Vandermeer, J. H. & Rosset, P. M.) 551–564 (McGraw Hill, New York, 1990).
49. Walker, B., Kinzig, A. & Langridge, J. Plant attribute diversity, resilience, and ecosystem function: the nature and significance of dominant and minor species. *Ecosystems* **2**, 95–113 (1999).
50. Burdon, J. J. The structure of pathogen populations in natural plant communities. *Annu. Rev. Phytopathol.* **31**, 305–323 (1993).
51. Wasilewska, L. Differences in development of soil nematode communities in single- and multi-species grass experimental treatments. *Appl. Soil Ecol.* **2**, 53–64 (1995).
52. Bertness, M. D. & Leonard, G. H. The role of positive interactions in communities: lessons from intertidal habitats. *Ecology* **78**, 1976–1989 (1997).
53. Nitta, T. Diversity of root fungal floras: its implications for soil-borne diseases and crop growth. *Jap. Agric. Res. Quart.* **25**, 6–11 (1991).
54. Elton, C. S. *The Ecology of Invasions by Animals and Plants* (Methuen, London, 1958).
55. Tilman, D. Community invasibility, recruitment limitation, and grassland biodiversity. *Ecology* **78**, 81–92 (1997).
56. Levine, J. M. & D'Antonio, C. M. Elton revisited: a review of evidence linking diversity and invasibility. *Oikos* **87**, 15–26 (1999).
57. Stohlgren, T. J. *et al.* Exotic plant species invade hot spots of native plant diversity. *Ecol. Monogr.* **69**, 25–46 (1999).
58. Lavorel, S., Prieur-Richard, A.-H. & Grigulis, K. Invasibility and diversity of plant communities: from patterns to processes. *Diversity Distrib.* **5**, 41–49 (1999).
59. McGrady-Steed, J., Harris, P. M. & Morin, P. J. Biodiversity regulates ecosystem predictability. *Nature* **390**, 162–165 (1997).
60. Orr, M. R. & Seike, S. H. Parasitoids deter foraging by Argentine ants (*Linepithema humile*) in their native habitat in Brazil. *Oecologia* **117**, 420–425 (1998).
61. Holway, D. A. Competitive mechanisms underlying the displacement of native ants by the invasive Argentine ant. *Ecology* **80**, 238–251 (1999).
62. Van Wilgen, B. W., Cowling, R. M. & Burgers, C. J. Valuation of ecosystem services: a case study from South African fynbos ecosystems. *BioScience* **46**, 184–189 (1996).
63. Zavaleta, E. S. in *Invasive Species in a Changing World* (eds Hobbs, R. J. & Mooney, H. A.) (Island, Washington DC, in the press).
64. Costanza, R. *et al.* The value of the world's ecosystem services and natural capital. *Nature* **387**, 253–260 (1997).
65. Stewart, G. & Hull, A. C. Cheatgrass (*Bromus tectorum* L.) — an ecological intruder in southern Idaho. *Ecology* **30**, 58–74 (1949).
66. Blockstein, D. E. Letter to the editor. *Science* **279**, 1831 (1998).
67. Naylor, R. L. Invasions in agriculture: assessing the cost of the golden apple snail in Asia. *Ambio* **25**, 443–448 (1996).
68. Porter, J. H., Parry, M. L. & Carter, T. R. The potential effects of climatic change on agricultural insect pests. *Agric. For. Meteorol.* **57**, 221–240 (1991).
69. Pimentel, D., Lach, L., Zuniga, R. & Morrison, D. Environmental and economic costs of nonindigenous species in the United States. *BioScience* **50**, 53–65 (2000).
70. USOT Assessment. *Harmful Non-indigenous Species in the United States* (US Government Printing Office, Washington, 1993).
71. Thorp, J. *The National Weeds Strategy: A Strategic Approach to Weed Problems of National Significance* (Agriculture and Resource Management Council of Australia and New Zealand, Canberra, 1997).
72. Jones, C. G., Ostfeld, R. S., Richard, M. P., Schaubert, E. M. & Wolff, J. O. Chain reactions linking acorns to gypsy moth outbreaks and Lyme disease risk. *Science* **279**, 1023–1026 (1998).
73. Fisher, A. C. & Hanemann, W. M. Option value and the extinction of species. *Adv. Appl. Micro-Econ.* **4**, 169–190 (1986).
74. Naylor, R. L. in *Invasive Species in a Changing World* (eds Hobbs, R. & Mooney, H. A.) (Island, Washington DC, in the press).
75. Channell, R. & Lomolino, M. V. Dynamic biogeography and conservation of endangered species. *Nature* **403**, 84–86 (2000).
76. Bryan, J. H., Foley, D. H. & Sutherland, R. W. Malaria transmission and climate change in Australia. *Med. J. Aust.* **164**, 345–347 (1996).
77. Morrison, J. I. *The Sustainable Use of Water in the Lower Colorado River Basin* (Pacific Institute and the Global Water Policy Project, Oakland, 1996).
78. Zavaleta, E. S. The emergence of waterfowl conservation among Yup'ik hunters in the Yukon-Kuskokwim Delta, Alaska. *Hum. Ecol.* **27**, 231–266 (2000).
79. Warren, D. M. & Pinkston, J. in *Linking Social and Ecological Systems* (eds Berkes, F. & Folke, C.) 158–189 (Cambridge Univ. Press, Cambridge, 1998).
80. Schlesinger, W. H. *et al.* Biological feedbacks in global desertification. *Science* **247**, 1043–1048 (1990).
81. Garrity, D. P. *et al.* The *Imperata* grasslands of tropical Asia: area, distribution, and typology. *Agrofor. Syst.* **36**, 1–29 (1997).
82. Peck, S. C. & Teisberg, T. J. in *Assessing Surprises and Nonlinearities in Greenhouse Warming* (eds Darmstadter, J. & Toman, M. A.) 80–83 (Resources for the Future, Washington, 1993).
83. Tomich, T. P., Kuusipalo, J., Menz, K. & Byron, N. *Imperata* economics and policy. *Agrofor. Syst.* **36**, 233–261 (1997).
84. Power, M. E., Matthews, W. J. & Steward, A. J. Grazing minnows, piscivorous bass, and stream algae: dynamics of a strong interaction. *Ecology* **66**, 1448–1456 (1985).
85. Johnson, K. G., Vogt, K. A., Clark, H. J., Schmitz, O. J. & Vogt, D. J. Biodiversity and the productivity and stability of ecosystems. *Trends Ecol. Evol.* **11**, 372–377 (1996).
86. Vitousek, P. M. & Hooper, D. U. in *Biodiversity and Ecosystem Function* (eds Schulze, E.-D. & Mooney, H. A.) 3–14 (Springer, Berlin, 1993).
87. Huston, M. A. Hidden treatments in ecological experiments: re-evaluating the ecosystem function of biodiversity. *Oecologia* **110**, 449–460 (1997).
88. Daily, G. C. *Nature's Services: Societal Dependence on Natural Ecosystems* (Island, Washington DC, 1997).
89. Goulder, L. H. & Kennedy, D. in *Nature's Services: Societal Dependence on Natural Ecosystems* (ed. Daily, G. C.) 23–48 (Island, Washington DC, 1997).

Acknowledgements

We thank B. R. Tershy for valuable inputs and J. D. Gerlach for access to his unpublished manuscript.

Systematic conservation planning

C. R. Margules* & R. L. Pressey†

*CSIRO Wildlife and Ecology, Tropical Forest Research Centre, and the Rainforest Cooperative Research Centre, PO Box 780, Atherton, Queensland 4883, Australia

†NSW National Parks and Wildlife Service, PO Box 402, Armidale, New South Wales 2350, Australia

The realization of conservation goals requires strategies for managing whole landscapes including areas allocated to both production and protection. Reserves alone are not adequate for nature conservation but they are the cornerstone on which regional strategies are built. Reserves have two main roles. They should sample or represent the biodiversity of each region and they should separate this biodiversity from processes that threaten its persistence. Existing reserve systems throughout the world contain a biased sample of biodiversity, usually that of remote places and other areas that are unsuitable for commercial activities. A more systematic approach to locating and designing reserves has been evolving and this approach will need to be implemented if a large proportion of today's biodiversity is to exist in a future of increasing numbers of people and their demands on natural resources.

It is an ancient and widespread human practice to set aside areas for the preservation of natural values. The sacred groves of Asia and Africa and royal hunting forests are historical examples^{1,2}. Other areas protect ecosystem services such as the delivery of clean water or the supply of timber, or mitigate the expected adverse effects of over-clearing³. Others protect recreational and scenic values and some have been planned to foster international cooperation⁴. Many of these areas meet the World Conservation Union's definition of a strictly protected area (IUCN categories I–IV)⁵, and hereafter we refer to such protected areas as 'reserves'. These areas are increasingly being complemented by reserves established principally for the protection of biodiversity, including ecosystems, biological assemblages, species and populations⁶. The basic role of reserves is to separate elements of biodiversity from processes that threaten their existence in the wild. They must do this within the constraints imposed by large and rapidly increasing numbers of humans in many parts of the world and their attendant requirements for space, materials and waste disposal⁷.

The extent to which reserves fulfil this role depends on how well they meet two objectives. The first is representativeness, a long-established goal referring to the need for reserves to represent, or sample, the full variety of biodiversity⁸, ideally at all levels of organization. The second is persistence. Reserves, once established, should promote the long-term survival of the species and other elements of biodiversity they contain by maintaining natural processes and viable populations and by excluding threats⁹. To meet these objectives, conservation planning must deal not only with the location of reserves in relation to natural physical and biological patterns but also with reserve design, which includes variables such as size, connectivity, replication, and alignment of boundaries, for example, with watersheds^{10,11}. A structured systematic approach to conservation planning provides the foundation needed to meet these objectives.

Systematic conservation planning has several distinctive characteristics. First, it requires clear choices about the features to be used as surrogates for overall biodiversity in the planning process. Second, it is based on explicit goals, preferably translated into quantitative, operational targets. Third, it recognizes the extent to which conservation goals

have been met in existing reserves. Fourth, it uses simple, explicit methods for locating and designing new reserves to complement existing ones in achieving goals. Fifth, it applies explicit criteria for implementing conservation action on the ground, especially with respect to the scheduling of protective management when not all candidate areas can be secured at once (usually). Sixth and finally, it adopts explicit objectives and mechanisms for maintaining the conditions within reserves that are required to foster the persistence of key natural features, together with monitoring of those features and adaptive management¹² as required. The effectiveness of systematic conservation planning comes from its efficiency in using limited resources to achieve conservation goals, its defensibility and flexibility in the face of competing land uses, and its accountability in allowing decisions to be critically reviewed. This is an idealized description of a process that is difficult to achieve in practice. Nevertheless, substantial parts have now been implemented around the world^{13–17} and some are used as illustrations below.

The practice of conservation planning has generally not been systematic and new reserves have often been located in places that do not contribute to the representation of biodiversity. The main reason is that reservation usually stops or slows the extraction of natural resources. In some regions, housing and commercial development compete with reserves for land¹⁸. The economic and political implications can be serious and reserves can be degraded or even lose their protected status when they prove to be economically valuable¹⁹. As a result, reserves tend to be concentrated on land that, at least at the time of establishment, was too remote or unproductive to be important economically²⁰. This means that many species occurring in productive landscapes or landscapes with development potential are not protected, even though disturbance, transformation to intensive uses, and fragmentation continue²¹. Another reason for the inappropriate location of reserves is the very diversity of reasons for which reserves are established. A diversity of goals means that different proponents see different places as important. Because highly valued areas arising from alternative conservation goals often fail to overlap²², there is competition among proponents for limited funds and the limited attention spans of decision-makers. Moreover, goals such as the protection of grand scenery and wilderness often focus on areas that are remote, rugged and

Figure 1 Social, economic and political factors often compete with reserves for land. **a**, Kings Canyon, Watarrka National Park, Northern Territory, Australia. This is a spectacular landscape, worthy of protection both for its outstanding natural beauty and for its biodiversity. But it is a remote and rugged area, valuable for tourism but not for extractive uses so it was easier to protect than more productive and economically valuable landscapes. **b**, An agricultural landscape in the Adelaide Hills, South Australia, with remnant woodland in the background. Remnants such as these contain species that are not represented in more remote and inaccessible areas, so their contribution to the overall goal of maintaining biodiversity is just as great. Despite their natural values it is always a difficult social and political decision to protect them because they have economic value as well as biodiversity value. Photographs by Liz Poon.



residual from intensive uses, giving them a political advantage over goals such as representativeness, which focus also on disturbed, economically productive landscapes (Fig. 1).

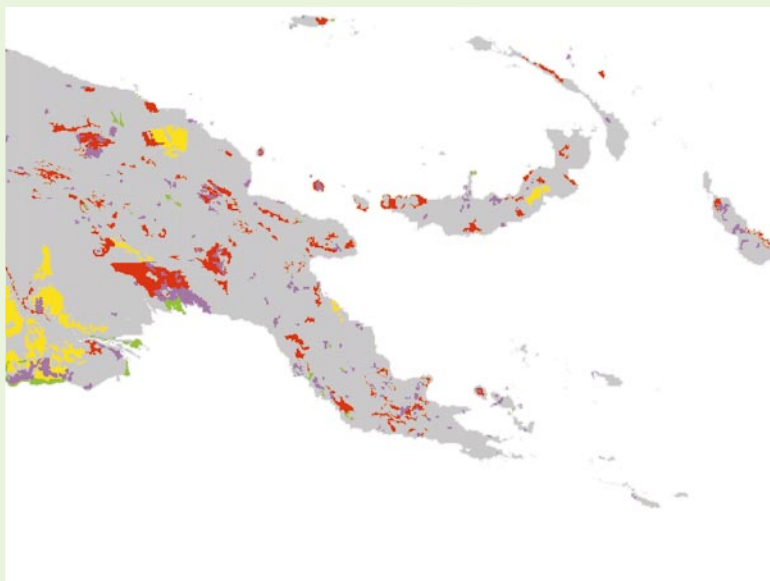
Conservation planning is therefore an activity in which social, economic and political imperatives modify, sometimes drastically, scientific prescriptions. This interaction need not be all one way. Science has at least three means of influencing the practice of nature conservation. First, an available body of scientific theory and application can provide some of the raw material for constructing policies²³. Second, science can offer solutions when called upon to assist in the implementation of policies and conventions, while also clarifying the social and economic implications of alternative methods or scenarios (this role is best filled when science is integral to the process, not simply called in for peer review²⁴ or when technical or political problems emerge). Third, science can and should be used to review the effectiveness of political processes for achieving stated biodiversity goals. A structured framework for conservation

planning will enhance the effectiveness with which science can do these three things.

A framework for systematic conservation planning

Systematic conservation planning can be seen as a process in six stages²⁵ (Box 1), each of which is discussed below with examples of the tasks and decisions required. The process is not unidirectional — there will be many feedbacks and reasons for revised decisions about priority areas. For example, it will be necessary to re-examine conservation goals as knowledge accumulates, and replacement candidate reserves will have to be identified when unforeseen difficulties arise in implementation. Although our discussion focuses on reserves, the framework applies equally well to many problems in 'off-reserve' conservation, including habitat restoration^{25,26}. Decisions about both on- and off-reserve conservation, if they are not to be *ad hoc* and uncoordinated, should be guided by explicit goals, identification of priorities in regional or broader contexts, and clear choices between

Figure 2 A map of biodiversity priority areas in Papua New Guinea¹⁶. The targets that are met by this set of areas are the representation of 608 environmental domains³⁷, 564 vegetation types, 10 species assemblages and 12 rare and threatened species. For the derivation of these targets, see refs 16, 92. In meeting targets, the set of areas also minimizes foregone opportunities for timber extraction, represents all existing reserves, minimizes the number of areas currently used for intensive agriculture, gives preference to areas with low human population density and gives preference to areas identified previously by experts as biodiversity priority areas⁹². The selected areas occupy 16.8% of the country and are inhabited by 210,000 people out of a population of approximately 4 million. A total of 398 areas were selected from 4,470 candidate areas or planning units. These units were aerial photograph patterns that were previously mapped for a database on agricultural and forestry suitability. The trade-off between biodiversity gain and opportunity costs, and the application of the other spatial constraints, was achieved with the TARGET software^{94,109}. The colours represent different index classes of timber volume. Yellow is highest, red next highest, purple next and green lowest.



potential conservation areas and alternative forms of management.

Stage 1. Measure and map biodiversity

Because of the complexity of biodiversity, surrogates such as sub-sets of species, species assemblages and habitat types have to be used as measures of biodiversity, and the locations of these surrogates within areas have to be plotted so that similarities or differences among areas can be estimated.

Biological systems are organized hierarchically from the molecular to the ecosystem level. Logical classes such as individuals, populations, species, communities and ecosystems are heterogeneous. Each member of each class can be distinguished from every other member. It is not even possible to enumerate all of the species of any one area, let alone the members of logical classes at lower levels such as populations and individuals. Yet this is biodiversity, and maintaining that complexity is the goal of conservation planning. For the foreseeable future it will be necessary to accept this incomplete knowledge and adopt methods for making the most of what we do know or can discover from new surveys. Thus, surrogate or partial measures of biodiversity must be used to estimate similarity or difference among areas within planning regions.

The choice of surrogate measures is not trivial. The strong temptation is to use a group of species: for example, vascular plants,

vertebrates or butterflies. We may know that the presence of a butterfly indicates the presence of its food plant somewhere nearby. The real question, however, is whether the presence of that butterfly, or any other taxon, indicates the presence of other taxa to the extent that it can be considered a suitable surrogate for overall biodiversity. Tests of taxonomic surrogacy in Britain²⁷ and South Africa²⁸ are not encouraging, but more promising results have been obtained in Uganda²⁹. Divergent results are attributable to differences in analytical methods, geographical scales and biogeographical histories of the study areas. Reliable generalizations and an understanding of how such factors affect taxonomic surrogacy are still developing. Higher levels in the biological hierarchy, such as species assemblages, habitat types and ecosystems lose biological precision, but have other advantages. They can integrate more of the ecological processes that contribute to the maintenance of ecosystem function³⁰ (although there is active debate on this issue³¹) and the relevant data are more widely and consistently available. In addition, there are sound theoretical reasons why environmental variables should be good estimators of the spatial distribution patterns of species^{32–34} and there are now some empirical studies that add support^{35–37}. New statistical techniques are also being developed to compare how well different environmental surrogates reflect the distribution patterns of species³⁸.

Box 1

Stages in systematic conservation planning

Systematic conservation planning can be separated into six stages, and some examples of tasks and decisions in each are presented below²⁵. Note that the process is not unidirectional; there will be many feedbacks and reasons for altering decisions (see text for examples).

1. Compile data on the biodiversity of the planning region

- Review existing data and decide on which data sets are sufficiently consistent to serve as surrogates for biodiversity across the planning region.
- If time allows, collect new data to augment or replace some existing data sets.
- Collect information on the localities of species considered to be rare and/or threatened in the region (these are likely to be missed or under-represented in conservation areas selected only on the basis of land classes such as vegetation types).

2. Identify conservation goals for the planning region

- Set quantitative conservation targets for species, vegetation types or other features (for example, at least three occurrences of each species, 1,500 ha of each vegetation type, or specific targets tailored to the conservation needs of individual features). Despite inevitable subjectivity in their formulation, the value of such goals is their explicitness.
- Set quantitative targets for minimum size, connectivity or other design criteria.
- Identify qualitative targets or preferences (for example, as far as possible, new conservation areas should have minimal previous disturbance from grazing or logging).

3. Review existing conservation areas

- Measure the extent to which quantitative targets for representation and design have been achieved by existing conservation areas.
- Identify the imminence of threat to under-represented features such as species or vegetation types, and the threats posed to areas that will be important in securing satisfactory design targets.

4. Select additional conservation areas

- Regard established conservation areas as 'constraints' or focal points for the design of an expanded system.
- Identify preliminary sets of new conservation areas for consideration as additions to established areas. Options for doing this include reserve selection algorithms or decision-support software to allow stakeholders to design expanded systems that achieve regional conservation goals subject to constraints such as existing reserves, acquisition budgets, or limits on feasible opportunity costs for other land uses.

5. Implement conservation actions

- Decide on the most appropriate or feasible form of management to be applied to individual areas (some management approaches will be fallbacks from the preferred option).
- If one or more selected areas prove to be unexpectedly degraded or difficult to protect, return to stage 4 and look for alternatives.
- Decide on the relative timing of conservation management when resources are insufficient to implement the whole system in the short term (usually).

6. Maintain the required values of conservation areas

- Set conservation goals at the level of individual conservation areas (for example, maintain seral habitats for one or more species for which the area is important). Ideally, these goals will acknowledge the particular values of the area in the context of the whole system.
- Implement management actions and zonings in and around each area to achieve the goals.
- Monitor key indicators that will reflect the success of management actions or zonings in achieving goals. Modify management as required.



Figure 3 White Rhinos currently persist in relatively small intensively managed populations in game reserves. Off-reserve management in suitable habitat would probably be necessary if populations were to return to self-sustaining levels, although conflict with human populations makes it extremely unlikely that this would ever happen. Photograph by Liz Poon.

Planning is essentially a matter of comparison so it is preferable to compare two or more areas with the same kind of information at the same level of detail. A map of vegetation types (communities or habitat types) and/or environmental classes provides spatial consistency across wide areas. On the other hand, museum and herbarium data on the locations of taxa are notoriously biased, having been collected for a different purpose (systematics), and often in an opportunistic manner, from the places that collectors expected to find what they were looking for or that were conveniently accessible^{39,40}. Plots of the field records from many collections therefore map road networks. Various methods — empirical, statistical and computational — are now available for modelling wider spatial distribution patterns from the point records that field samples represent^{41–43}, but their reliability is also at least partly a function of the degree of spatial bias. New systematic field surveys to fill gaps are the best solution but they can be expensive and time consuming.

There is no best surrogate. The decision on which to use will depend on many factors including what data are available and what resources there are for data analysis (for example, spatial modelling) and the collection of new data. In most parts of the world, the only spatially consistent information available is on higher-order surrogates such as vegetation types and environmental classes. Collections of taxa might form an accurate representation of some biological distributions in some countries where well designed and well resourced surveys have been used to collect the data. Taxa collections may also be used with some reliability at coarse scales (for example, grid cells of 50 km × 50 km), but usually become less reliable at the scale of individual reserves⁴⁴. If taxa sub-sets are used without spatial modelling, it is usually with the understanding that the disadvantage of spatial bias is offset by the advantage of having at least some direct biological information to complement higher-order surrogates. Combinations of surrogates will be most practicable in most situations. In a recent study in Papua New Guinea, environmental domains classified from climate, landform and geology³⁷, vegetation types mapped from aerial photographs, and the known locations of rare and threatened species were all used as biodiversity surrogates (Fig. 2)¹⁶.

A decision is also needed at this stage on how to define planning units, the building blocks of the reserve system. Planning units can be regular (for example, grids or hexagons) or irregular (for example, tenure parcels, watersheds or habitat remnants). A mix of planning units might be appropriate in regions that contain both fragmented landscapes and extensive tracts of uncleared vegetation. The choice has implications for the efficiency with which representation goals can be achieved as well as for the design and management of reserves⁴⁵. For the reserve selection process described in stage 4, it is

necessary to compile data on biodiversity surrogates for each of the planning units in the region. Data on tenure (for stages 3, 4 and 5, below) and other contextual data that might influence selection and implementation (for example, roads, rivers, terrain, timber resources and threats) should also be compiled at this stage.

Stage 2. Identify conservation goals for the planning region

The overall goals of systematic conservation planning — representativeness and persistence — have to be translated into more specific, preferably quantitative, targets for operational use. Targets allow clear identification of the contributions of existing reserves to regional goals and provide the means for measuring the conservation value of different areas during the area selection process in stage 4 below. Targets such as 10 or 12% of the areas of countries or vegetation types have been criticized because they are too small to prevent the extinction of many species, can be subverted by reserving the least productive and least threatened landscapes, and can mislead the public into believing that limited conservation action is adequate⁴⁶. A focus on targets for reserves may also remove incentives to implement other conservation actions such as off-reserve management¹. These criticisms are valid, but are aimed at how targets are set rather than exposing reasons for not setting targets at all. Planners need to know what they are aiming for. 'More equals better' is good in principle, but does little to resolve choices between areas with different biotas when other demands narrow the geographical scope for reservation. Accordingly, planners need targets that do several things: focus on scales that are much finer than whole countries or regions; deal with natural processes as well as biodiversity pattern; reflect the relative needs of species and landscapes for protection; recognize that reserves must be complemented by off-reserve management, preferably also with targets; and leave options open for revision as social and economic conditions change. Ideally, reservation targets will be an integral part of policies and government processes⁴⁷. Failure to achieve targets for economically valuable landscapes is likely, so periodic reviews (stage 3, below) are necessary.

Most exercises in systematic conservation planning have chosen areas on the basis of the occurrences of species. Some have used predicted probabilities of occurrence⁴⁸. Recent applications have set targets for the spatial extent of communities, habitat types or environmental classes, sometimes with explicit formulae for adjusting targets according to factors such as natural rarity and vulnerability to threats¹³. These are all targets for representing a biodiversity pattern. Targets for ecological processes can be more problematic. Because conservation planning is a spatial exercise, protection of natural processes must be based on their spatial surrogates rather than the processes themselves (for example, size, lack of roads, watershed

boundaries, and migration routes). Setting process targets can be difficult in practice because the environment is heterogeneous in space and time and different species function at different spatial and temporal scales⁴⁹. Nevertheless, seven aspects of theory on ecological and evolutionary processes, now supported by some empirical evidence, can provide guidelines.

Biogeographical theory

Traditionally, the equilibrium theory of island biogeography⁵⁰ and associated biogeographical theory has been used to help set targets for size, shape and distance between reserves (although usually such targets were not quantitative). This body of theory tells us that bigger reserves are better, the closer they are the better, the more circular the better, and that reserves should be linked by habitat corridors^{51,52}. In the real world of conservation planning, the opportunity to apply such guidelines is constrained by costs and patterns of land-use history. These design principles also introduced an important trade-off into planning that is seldom acknowledged. If the area available for reservation is limited, a choice might have to be made between a few large reserves that favour the persistence of some species or more smaller reserves that together are more representative of the region's biodiversity but individually are less effective for the persistence of some species, for example, large, wide-ranging species^{17,53}. An early and widely ignored criticism of the equilibrium theory was that it treated islands as featureless plains with no internal habitat diversity and species as characterless features with no genetic or geographical variation⁵⁴. There is now some experimental support for the prediction that increased isolation reduces the likelihood of persistence of certain species⁵⁵, supporting targets for connectivity. However, attention has rightly shifted to the roles of environmental heterogeneity, species interactions, local- and regional-scale population dynamics, and the effects of habitat modification in reserve planning.

Metapopulation dynamics

In general, a metapopulation⁵⁶ is a network of local populations linked by dispersal. More narrowly, the term is used to describe systems in which local populations periodically go extinct with recolonization occurring by migration from other local populations⁵⁷. Metapopulations go extinct when the rate of extinction of local populations exceeds the rate of migration and recolonization. Confining a species to a reserve may disrupt metapopulation dynamics, increasing the risk of local extinction due, for example, to a catastrophic event such as wildfire, and decreasing the chances of recolonization. Metapopulation theory calls for targets that consider reservation across species' natural ranges so that some populations might escape the impact of unpredictable events, thereby spreading the risk of extinction⁵⁸. It also calls for the retention of landscape linkages to promote dispersal and the exchange of individuals between geographically separate sub-populations⁵⁹ and for

the retention of patches of suitable, but currently unoccupied, habitat⁶⁰.

Source-pool effects and successional pathways

The species composition of an area changes over time in a process usually called ecological succession. Some of these changes will be due to dispersal but others will be the products of initial conditions. There is a mix of starting propagules available in an area and subsequent changes reflect a sorting of this mix according to life-history traits and interspecific interactions⁶¹. Because of periodic, patchy disturbances, most regions contain areas at various stages along these pathways and many species exploit the temporal and spatial variation of natural disturbance regimes⁶². The implications for target setting are that all successional stages might need to be represented, replication of reserves to sample different successional stages might be desirable, and large reserves are better because they can better accommodate natural patch dynamics without succession being reset throughout by a single event such as a wildfire⁶³.

Spatial autecological requirements

Different species require different amounts of space to complete their life cycles⁵⁷ (Fig. 3). Most reserves contain one or more species that would not persist as residents even for one generation if they became isolated. Many other reserves, without supplementation by unreserved habitat, would be likely to lose species in the long term through a variety of chance events. Thus, the long-term persistence of some taxa requires sustainable populations across entire landscapes or regions as predicted, for example, for the northern spotted owl (*Strix occidentalis caurina*) in the Pacific northwest United States⁶⁴. There is a vast literature on population viability analysis^{65,66}. Reservation targets should include viable population sizes and structures (for example, age classes and sex ratios) when these are known. Many species exploit temporal variation by moving between different habitats, requiring targets to recognize key habitat combinations where these can be identified. The focal species approach⁶⁷ attempts to integrate patterns and processes by identifying those species in a landscape that are most demanding of resources and then targeting them for management. The kinds of resources needed by focal species may be, for example, large areas, connectivity between habitat patches and complex heterogeneous habitats¹⁷. The argument is that if management can maintain these species in a landscape, then most other species will be maintained as well.

Source-sink population structures

If, in some high-quality habitats (sources) a species' reproduction rate exceeds mortality, but in low-quality habitats (sinks) its reproduction rate is lower than mortality, then a net dispersal away from sources may sustain populations in sinks^{57,68}. In southeastern Australia, 63% of the arboreal marsupial population is found in only 9% of the forest with high foliar nutrients⁶⁹. Dispersal throughout the remainder of the forest occurs from these areas of high population



Figure 4 Isolated habitat remnants in the wheat belt of Western Australia. Isolation causes physical changes to habitat remnants, which in turn can lead to changes in species composition and population sizes. Photograph courtesy of CSIRO, Wildlife & Ecology.

density. If population sources for some species are outside reserves or are not targeted for reservation, then the presence of those species within reserves is at risk.

Effects of habitat modification

If reserves become remnants of natural habitat surrounded by alien habitat such as cropland or pasture, changes brought about by isolation and exposure have implications for the persistence of species within them (Fig. 4). Changes in fluxes of wind, water and solar radiation⁷⁰ can lead, in turn, to changes in vegetation structure, microclimate, ground cover and nutrient status⁷¹. These changes may favour some species, but they also lead to reduced population sizes and local extinction of others^{72,73}. Once isolated and exposed, habitat remnants may be placed on a trajectory of continued change. Deleterious effects can feed back on themselves to increase their magnitude⁷⁴, they can simply accumulate with time⁷⁵, or they can cascade, with a change in a species' abundance or productivity leading to unforeseen changes in the populations of other species. In

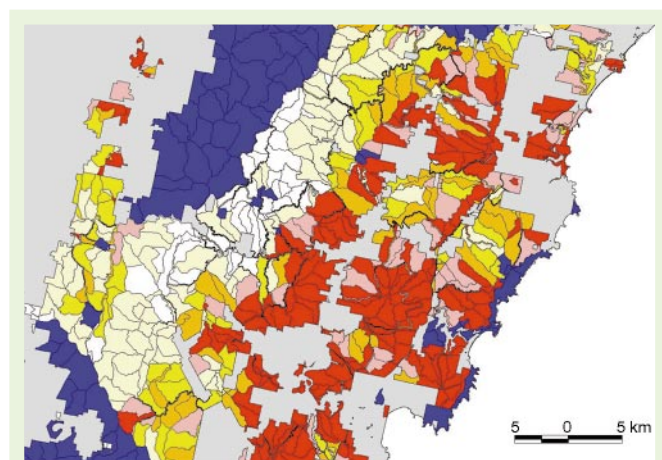


Figure 5 Pattern of complementarity on part of the south coast of New South Wales. The map is based on the same data used in the C-Plan decision-support system¹³ in late 1999 to guide negotiations between interest groups over new forest reserves in the region. The eastern boundary is the coastline. Blue areas are reserves established before the negotiations. Grey areas are tenures not considered in the planning process. Other polygons are logging compartments (average area about 200 ha or 2×10^6 m²) used as the building blocks of the expanded reserve system. Colours of these indicate five intervals of 'percentage contribution', the measure used in this exercise to indicate complementarity with existing reserves. Highest values are red (81–100%) and grade through pink, orange, dark yellow, pale yellow (>0–20%) and white (0%). Values of percentage contribution are based on reservation targets (in hectares) for each of 107 forest ecosystems in the region. Percentage contribution is calculated in two stages. In the first stage, a contribution value (in hectares) for each forest ecosystem in each compartment is calculated using two rules — if $A_i \leq T_i$ then $C_i = A_i$; if $A_i > T_i$ then $C_i = T_i$, where A_i is the extent of forest ecosystem i in the compartment, T_i is the remaining regional reservation target for the forest ecosystem, taking into account the contributions of existing reserves and any compartments previously given notional reserve status, and C_i is the current contribution of the compartment's sample of the forest ecosystem to the target. In the second stage, percentage contribution of the compartment is calculated as the sum of C values across all the forest ecosystems it contains, expressed as a percentage of the compartment's area. Compartments with highest values are largely or fully occupied by forest ecosystems well below target. Compartments with zero values contain only forest ecosystems with targets already achieved. Complementarity values show a marked association with distance eastwards from the large reserves in the westerly escarpment and more rugged foothills. In contrast, the small coastal reserves have no apparent influence on the complementarity of the adjacent compartments because they contain little forest. In the far northwest, higher values reflect the occurrence of forest ecosystems of the tableland, which are poorly reserved in the nearby escarpment reserves. Because complementarity is dynamic, percentage contribution was recalculated and redisplayed during the negotiations whenever one or more compartments were notionally reserved.

fragmented landscapes, where reserves are likely to be small and isolated, targets for off-reserve conservation are particularly important and they should include buffers around remnants, sympathetic management of poorly protected vegetation types or environments, and habitat restoration

Species as evolutionary units

It has long been argued that species should be treated as dynamic evolutionary units rather than as types^{76,77}. There are at least two related planning implications. First, areas occupied by taxa that appear from phylogenies to be actively radiating, or are most phylogenetically distinct, might be targeted for protection^{78,79}. Second, with an understanding of the physical and biological processes leading to active diversification of taxa, it is possible to identify and set targets for evolutionary templates. The most distinctive evolutionary feature of the Cape Floristic Region of South Africa has been the recent and massive diversification of many plant lineages. This process has been related to landscape features, such as interfaces between different soil types, which are now targeted for conservation action¹⁴.

These seven aspects of ecological and evolutionary processes have been explored largely as independent lines of research, separate also from the extensive work on the derivation and application of targets for biological patterns. An integration of all these research areas is needed for planning applications if the goals of representation and persistence are to be achieved⁴⁹. The best way forward is not yet clear but some attempts in different regions by different planning groups will allow comparisons to be made and, hopefully, some promising directions to be identified.

Stage 3. Review existing reserves

The extent to which targets for representation and persistence have already been achieved in existing reserves has to be determined. This defines the scope of the task in stage 4. Systematic reviews of existing reserve systems have a long history and are the conceptual basis for the Gap Analysis Program in the United States, now incorporating research and development projects and their applications in the 48 contiguous states⁴⁴. This programme was designed originally to identify gaps in the coverage of reserve networks but its increasing activity in identifying candidate conservation areas⁸⁰ (stage 4, below) is grounded in the systematic planning methods described here, which from the earliest applications have recognized the contribution of existing reserves to explicit targets⁸¹.

Analyses of gaps in networks of reserves have concentrated on which features are represented or not represented and to what extent. Two other aspects of gap analysis have received little attention. The first is the relative imminence or likelihood of species or habitats becoming extinct without conservation action. Because features that are under-reserved according to representation targets vary in their exposure and vulnerability to threatening processes, some gaps are more important than others⁸². Decisions about the scheduling of conservation action relative to threat are crucial for effective implementation (stage 5, below). Gap analyses that incorporate threats can reveal spatial biases in action by agencies and governments that inhibit effective implementation.

The second neglected aspect of gap analysis relates to natural dynamics and the persistence of biodiversity in the long term. Measures of gaps in process and persistence are few^{11,83} and a comprehensive, generic set of criteria for measuring gaps in the coverage of processes is lacking. Although most planners would agree that large size, connectivity and integrity are generally desirable, many species and vegetation types now exist only in remnants of habitat that are altered and surrounded by intensive land uses. The criteria for assessing gaps in coverage will be different in fragmented landscapes than in landscapes in which large contiguous tracts of habitat remain. The relative priority of reserve design criteria when they produce contrasting results (for example, compactness versus replication) has not been adequately addressed, nor has the role of partial contributions to biodiversity protection from areas under different

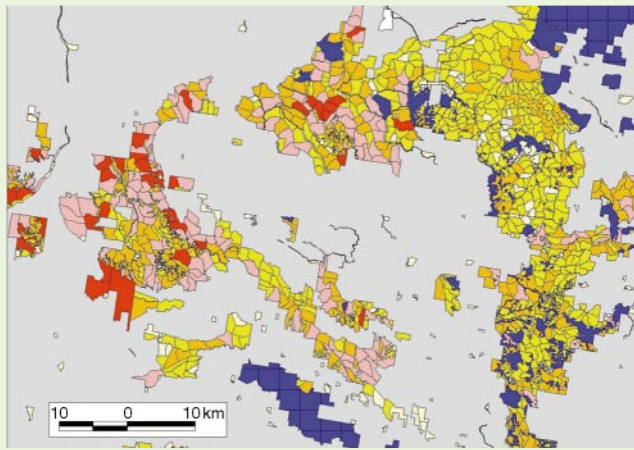


Figure 6 Pattern of irreplaceability in part of the northeast forests of New South Wales. The map is based on the same data and calculations of irreplaceability used in the C-Plan decision-support system¹³ in 1998 to guide negotiations between interest groups over new reserves in the region. Blue areas, grey areas and other polygons as in Fig. 5. The gradient from red to white indicates irreplaceability values of logging compartments based on the mix of forest ecosystems within each compartment, the distributions of 198 forest ecosystems across the region, their individual reservation targets in hectares, and the extent to which each target is already met in the existing reserves. Red areas are totally irreplaceable; if they are not reserved, one or more targets will not be met. Progressively lower values (pink, orange, dark yellow, pale yellow and white) indicate logging compartments with progressively more replacements. With lower values, the options for achieving targets are less constrained if compartments are unavailable or prove unsuitable for reservation. Like complementarity, irreplaceability is a dynamic measure. In the 1998 negotiations, values were recalculated each time one or more compartments were notionally reserved.

management regimes outside strict reserves.

In most planning exercises, implementation (stage 5, below) is likely to be gradual or, if rapid, will often fail to achieve all targets, particularly those for landscapes with economic potential. In these cases, the planning process should loop back periodically from stage 5 to stage 3 so that progress can be updated, new areas selected as appropriate (stage 4), and implementation reconsidered.

Stage 4. Select additional reserves

After the review of existing reserves, the need for additional areas to achieve the outstanding targets will become clear. At least some of the area selections at this stage are only preliminary because implementation (stage 5) invariably reveals practical impediments that require a degree of revision of the initial choices. The existing reserves are recognized not only for their contributions to targets but also because they can become the focal points or spatial constraints around which enlarged reserves or new, separate ones are located. The most convenient tools for the task of selection are algorithms, which apply explicit rules to identify notional sets of areas⁸⁴. These algorithms can be used to investigate various policy options, for example, to include or exclude wilderness areas, old-growth forest or regenerating areas, and to compare outcomes in terms of the number or total extent of new reserves needed. They can also indicate to planners whether the full set of targets is achievable within the expected limits of land area, acquisition cost or opportunity costs for other uses and, if not all are possible, the extent to which trade-offs are necessary (for example, between efficiency and design, or between representation of all forest types and the requirements of industry for timber). They provide a basis for negotiation or refinement of the conservation plan by regional or local experts. A recent development is the incorporation of algorithms into decision-support systems to guide structured negotiations between interest groups¹³. Used in this way, algorithms are able to guide decisions not only about how reserves sample biodiversity, but also about the design of reserve systems.

Complementarity

All selection algorithms use complementarity, a measure of the extent to which an area, or set of areas, contributes unrepresented features to an existing area or set of areas^{78,85}. The precise measure depends on the targets that have been identified and on the type of data. Most simply, it can be thought of as the number of unrepresented species (or other biodiversity features) that a new area adds. It has also been interpreted as a similarity index based on the number of species shared and not shared between two areas^{29,86}, as the contribution a new area makes to sampling a complete multivariate pattern generated by a classification or ordination of all areas⁸⁷, and as the distance in multivariate space that a new area is from existing areas^{88,89}. For targets set in terms of the extent of features such as forest types, complementarity can be measured as the contribution an area makes to outstanding targets according to the proportions of different types within that area (Fig. 5). An area with high complementarity will not necessarily be the richest⁹⁰. If, for example, an area contributes few species or habitat types and those features are not widely represented in the landscape, then its complementarity value could be extremely high. Another important property of complementarity is that it is recalculated for all unselected areas each time a new area is added to the notional reserved set. This recognizes that the potential contribution of an area to a set of targets is dynamic — some or all of the features in an unselected area might have had their targets partly or fully met by the selection of other areas. In contrast, more traditional measures of conservation value such as species richness or the number of rare species are unresponsive to changing targets and decisions to reserve other areas.

Spatial constraints on the selection of reserves

Constraints on the area selection process can be grouped into five kinds. The first, irreplaceability⁹¹, is inherent in any data set. When selection algorithms or regional experts decide on areas for reservation they choose between alternative areas for meeting conservation targets. For some planning exercises, it can be useful to display these alternatives explicitly as a map of irreplaceability (Fig. 6), indicating for each of the areas in a region the options for replacing it while still achieving conservation targets. Some areas have no replacements, whereas others have many. This information can be used to indicate the scope for altering selections by algorithms or experts (for example in trade-offs between targets and extractive land uses), to guide negotiations over new conservation areas, or to set priorities for implementation (stage 5, below). Four other spatial constraints are described below with examples from a recent application in Papua New Guinea (PNG) (Fig. 2)^{16,92}.

Costs. The use of an area for the protection of biodiversity generally means that it should not be available for commercial uses. Thus, biodiversity protection incurs opportunity costs. Trade-offs between opportunity costs and biodiversity gain can be achieved during the area selection process⁹³ or as a separate exercise after an initial selection¹³. It is important for the credibility of conservation planning that conservation goals are seen to be achieved in a way that minimizes, as far as possible, forgone opportunities for production. It is now possible to measure the opportunity costs of achieving a biodiversity goal and, conversely, the biodiversity costs of meeting a production goal, where that goal requires land allocation^{94,95}. Examples of opportunity costs are timber volume and agricultural production. Figure 2 shows the relative timber volumes on selected biodiversity priority areas in PNG. Other kinds of costs such as acquisition costs and the ongoing costs associated with management and maintenance could also be incorporated as constraints in the area selection process.

Commitments. Commitments are areas that must be selected regardless of their contribution to targets. The most common examples are existing reserves (Fig. 5). Other examples might be areas containing rare and threatened species and areas of endemism. Both existing reserves and areas containing rare and threatened species were used in the PNG study. Existing reserves can also require additional commitments of areas, for example when they need to be linked or have

their boundaries rationalized.

Masks. These are areas to be excluded from selection. For the PNG plan, areas smaller than 10 km² and areas used intensively for agriculture were masked initially. It was found, however, that some areas heavily used for agriculture were required if the biodiversity goal was to be achieved because they represented environments that were unavailable for selection elsewhere¹⁶.

Preferences. Sometimes certain characteristics of areas for conservation are to be preferred, if possible, over others. For the PNG plan, areas with low human population density and areas previously identified by expert taxonomists and ecologists as biodiversity priorities were given preference for selection, where there was a choice. Combining expert assessments with explicit analyses of spatially consistent data has advantages. Experts are inevitably biased geographically and taxonomically. On the other hand, data matrices inevitably lack the full store of knowledge in experts' heads.

Stage 5. Implement conservation actions on the ground

There is a world of difference between the selection process described above, and making things happen on the ground. Implementation is usually complicated by the variety of people, agencies and commercial interests with a stake in the region and by the time needed to apply conservation management to particular areas. The eventual system of reserves can be very different from the one designed in stage 4.

An example of a relatively straightforward case of implementation is the 1996 expansion of forest conservation areas in eastern New South Wales, Australia¹³ (Fig. 7). Planning was restricted to public land and the application of conservation action was rapid once the new areas had been negotiated to meet (most) targets and boundaries had been fine-tuned on the ground. Only a few forms of protection were at issue with little uncertainty about where they should most appropriately be applied. The implemented configuration was little different from that produced in the selection stage. A more complex and probably more widespread situation involves a mix of land tenures, ongoing loss and alteration of indigenous vegetation during a protracted process of applying conservation action on the ground, and the need to decide on an appropriate mix of protection measures. Three types of decisions are particularly important⁹⁶. First, the most appropriate or feasible form of management should be identified for each area. This might be complicated by the need to apply particular forms of management in particular designated places, for example in biosphere reserves, which have core and buffer zones. In some cases, the preferred form of management might be infeasible and will need to be changed. Second, if one or more selected areas prove to be unexpectedly degraded or difficult to protect, it will be necessary to return to stage 4 and identify replacements, where they exist (Fig. 6). Third, decisions are needed on the relative timing of conservation action when resources are insufficient to implement the whole network quickly. With ongoing loss and alteration of habitat, a strategy is needed to minimize the extent to which conservation targets are compromised before being achieved.

One strategy for scheduling conservation action within regions is to plot selected areas on two axes⁹⁶. The first is irreplaceability or the extent to which the loss of the area will compromise regional conservation targets⁹¹. The second is vulnerability or the risk of the area being transformed by extractive uses. Areas with high values for both should receive priority for conservation action (Fig. 8). They are most likely to be lost and, because of the absence or small numbers of replacements, their loss will have the most serious impact on the achievement of targets. This approach is similar conceptually to the original definition of global hotspots^{97–99} and to other assessments of priorities at global or continental scales^{100–103}. Plotting selected areas on two axes also has an advantage over combining values for both to produce a single priority score. Different areas of the graph can indicate the need for alternative management prescriptions, subject to regular review (Fig. 8). Three important qualifications are necessary. One is that an exercise in triage¹⁰⁴ might be necessary to decide if strict

Figure 7 Spotted gum, *Eucalyptus maculata*, with an understorey of the cycad, *Macrozamia communis* in southeastern New South Wales, Australia. These forests are now the subject of a regional forest agreement, which allocates some areas to protection based on the contribution they make to agreed biodiversity targets, and allocates other areas to production based on agreed timber harvesting targets¹³.



Most public forests in eastern New South Wales and in other Australian states now have regional forest agreements in place. Photograph by Liz Poon.

reservation is infeasible for some very high priority areas, necessitating other forms of protection such as management agreements with landholders, or outright abandonment. A second is the unresolved question of whether and how vulnerability to different threatening processes (for example, clearing, logging and grazing) should be combined for prioritization. A third is that the idea has been used mainly to prioritize areas for achieving biodiversity pattern targets and has yet to be developed fully for process targets. For some process targets it will be necessary to combine individual candidate areas into larger units before identifying priorities. Conservation planners are then likely to confront some difficult choices. They will often have to decide whether a limited annual budget should be used, for example, to keep intact a movement corridor for ungulates, a block of habitat considered minimal for the viability of a carnivore species, or the only known location of an endemic plant¹⁴. Planning for both the representation of patterns and persistence of species and natural processes requires planners to compare apples and oranges. There are no guidelines for optimizing the outcome and no guarantees that the anticipated outcome will be realized.

Stage 6. Management and monitoring of reserves

Establishing a reserve heralds the beginning of another process that is at least as demanding as the preceding planning process and spans a much longer period of time. Management of reserves should ensure that their natural values are retained in the face of internal natural dynamics, disturbances from outside, and a variety of valid human uses. In practice, the management of many reserves is inadequately resourced, unplanned and often threatened by illegal use for basic human subsistence or commercial activities^{105,106}. Some exist only on paper, never having been implemented⁷.

Sound management effectively involves another cycle of the previous five stages applied to individual reserves. It requires information on the biodiversity of each reserve, knowledge of the processes that underpin ecological functions, and an understanding of the responses of key elements of biodiversity to natural processes and anthropogenic disturbances (stage 1). Management should be based on explicit goals or targets¹⁰⁷ (as in stage 2), preferably acknowledging the contribution of each reserve's particular natural values to the regional system. Based on the extent to which management goals are already being achieved (stage 3), it might be necessary to review prescriptions or zonings and to prepare a new management plan indicating which parts of reserves are appropriate for different uses, require regulation of natural processes or need to be rehabilitated (stage 4). Problems with implementation of the management plan (stage 5) will usually be minimized or avoided if key interest groups are consulted during its development. As with the selection and implementation of new reserves, this process is not fixed and

unidirectional. New data on patterns and processes within a reserve might call for revised goals. More generally, ongoing management should be complemented by periodic monitoring (back to stage 3) to assess the effectiveness of management actions in achieving nominated goals, with subsequent adjustment of goals and activities as appropriate. Adaptive management¹², coupled with a genuine commitment to monitoring, is increasingly recognized as crucial, not only to follow the status of selected elements of biodiversity, but also to assess the adequacy of resources for management, the capability of the responsible institutions, and the accountability with which funds are being used¹⁰⁸.

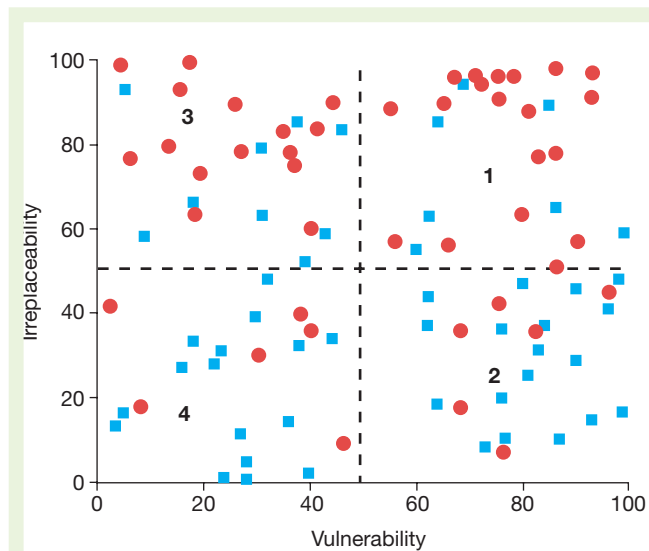


Figure 8 A framework for identifying priority conservation areas in time and space, applicable within regions to environments or other land types²⁰ or to potential conservation areas^{14,96}. The graph shows hypothetical data for 100 potential conservation areas, each with values for irreplaceability and vulnerability (for example, agricultural potential). Red points are a subset of areas that are notionally selected to achieve targets but not yet given reservation status. Blue points are possible replacements. Selected areas occur in all parts of the graph, although the selected proportion increases with higher irreplaceability. If not all selected areas can be protected immediately (a common situation), the positions of areas in the graph will change over time. Some of the more vulnerable areas are likely to be converted to agriculture. As this happens, the irreplaceability of some of the remaining areas will increase as they become more important for achieving targets for features that are now less extensive and/or less frequent elsewhere. Conversely, as areas are progressively reserved, the irreplaceability of others will decrease as the features they contain approach or reach their conservation targets. The vulnerability of areas will also change, most likely upwards as land-use pressures intensify. Appropriate responses by conservation planners can be related to the different quadrants, as in previous studies at broader scales^{100–102}. Quadrant 1: areas most likely to be lost and with fewest replacements. Protection is urgent if targets are not to be compromised. Some will probably be fragments of previously extensive vegetation types where strict reservation is difficult to apply (private tenure) or impractical (management liabilities) and must be supplemented with off-reserve management. Quadrant 2: areas vulnerable to loss but with more replacements, either because features are relatively common and extensive relative to targets or because targets have been partly met in existing reserves. Holding measures are necessary to avoid loss of some areas causing others to move upwards into quadrant 1. Options for protection include reservation where appropriate (and without pre-empting reserves in quadrant 1) complemented with off-reserve management. Quadrant 3: areas with lower present risk of agricultural conversion but high irreplaceability (for example, rocky ranges in a matrix of agricultural land or rare land types outside the climatic limits of agriculture). Protection is less urgent and acquisition for reservation more feasible than for quadrants 1 and 2 because of slower rates of transformation and (likely) lower land prices. Quadrant 4: the positions of areas here are likely to be stable and require least intervention, although monitoring of land use is advisable.

Interaction between reserve management and the location and design of reserves is inevitable. Decisions in the earlier stages of the planning process should, if possible, anticipate management issues. Key considerations include size and shape, alignment of boundaries with watersheds, avoidance of intrusive adjacent land uses, negotiations with neighbours, and the maintenance of migration routes. In turn, as the management needs of established reserves become apparent or as new needs emerge, it might be necessary to return to the selection stage (stage 4) to modify the design of individual reserves or the overall conservation network.

The outlook

There are many views about how best to identify priority conservation areas. To some extent this diversity is welcome as it arises from attempts by people with varying backgrounds to solve different problems in different parts of the world. This variety contributes usefully to an ongoing debate about appropriate planning approaches. But some of the divergence is less useful and seems to reflect different, poorly defined conservation goals and different, often implicit assumptions about the constraints under which conservation action will be applied. If these goals and assumptions were defined more explicitly, the relative roles and limitations of alternative approaches might be better understood and more attention could focus on addressing particular knowledge gaps and problems of implementation and management. Both clarity of purpose and productive debate would be achieved more readily if there was more direct interchange between groups working on conservation planning and between these groups and managers who face the daily challenges of staving off threats to biodiversity. Better communication depends inevitably on the interactions between individual researchers and managers, but more regular organized meetings specifically for conservation planning could also achieve much.

Conservation planning is also riddled with uncertainty. In the six stages of planning described here, uncertainty pervades the use of biodiversity surrogates, the setting of conservation targets, decisions about which kinds of land tenure can be expected to contribute to targets and for which features, and decisions about how best to locate, design, implement and manage new conservation areas in the face of limited resources, competition for other uses, and incursions from surrounding areas. New developments in all the planning stages will progressively reduce, but never eliminate, these uncertainties. One implication is that planners, rather than proceeding as if certain, must learn to deal explicitly with uncertainty in ways that minimize the chances of serious mistakes.

An urgent need is for more precision in the measurement of biodiversity and more consistency in mapping it across regions and biomes. In part this can be addressed by the allocation of more resources — funds, personnel and infrastructure — to the collection of field records of species and other biodiversity features. However, because all collections are samples, and as complete inventories of regions are not a realistic option for the foreseeable future, the design of data-collecting activities should be based soundly in ecological theory and should enable the application of proven statistical techniques to the modelling of wider spatial distribution patterns from the point locations that these field records were taken from³⁹.

Another need is for more effort to be applied to mapping patterns and monitoring rates of spread of threats to biodiversity, as it is such threats to which conservation planning should respond. A better understanding of the present and future distribution patterns of various threats will help focus limited conservation resources on areas and features most at risk. It will also clarify the extent to which conservation priorities overlap with priority areas for extractive and destructive uses. Some threats arise for reasons that can be understood only with the benefit of hindsight, but this is no reason not to improve foresight with refined predictions about the effects of extractive uses, urbanization and the spread of alien species.

More precise management prescriptions for the persistence of

biodiversity are also needed. So far, enough is known only about a select few species, mostly large vertebrates and vascular plants, for effective management prescriptions. Finally, and just as importantly, biologists and ecologists must participate more in real planning processes. This is the only sure way to understand fully where the need for new ecological and biological knowledge is and what the social and political constraints on effective planning really are. □

- Kanowski, P. J., Gilmour, D. A., Margules, C. R. & Potter, C. S. *International Forest Conservation: Protected Areas and Beyond* (Commonwealth of Australia, Canberra, 1999).
- Chandrashekar, U. M. & Sankar, S. Ecology and management of sacred groves in Kerala, India. *For. Ecol. Mgmt* **112**, 165–177 (1998).
- Grove, R. H. Origins of western environmentalism. *Sci. Am.* **267**, 22–27 (1992).
- Hanks, J. Protected areas during and after conflict: the objectives and activities of the Peace Parks Foundation. *Parks* **7**, 11–24 (1997).
- World Conservation Union. *Guidelines for Protected Area Management Categories* (IUCN, Gland, Switzerland, and Cambridge, 1994).
- Anon. *Global Biodiversity Strategy* (World Resources Institute, World Conservation Union, and United Nations Development Program, Washington DC, 1992).
- Terborgh, J. *Requiem for Nature* (Island, Washington DC, 1999).
- Austin, M. P. & Margules, C. R. in *Wildlife Conservation Evaluation* (ed. Usher, M. B.) 45–67 (Chapman & Hall, London, 1986).
- Soulé, M. E. (ed.) *Viable Populations for Conservation* (Cambridge Univ. Press, Cambridge, 1987).
- Shafer, C. L. National park and reserve planning to protect biological diversity: some basic elements. *Landscape Urban Plan.* **44**, 123–153 (1999).
- Peres, C. A. & Terborgh, J. W. Amazonian nature reserves: an analysis of the defensibility status of existing conservation units and design criteria for the future. *Conserv. Biol.* **9**, 34–46 (1995).
- Holling, C. S. (ed.) *Adaptive Environmental Assessment and Management* (International Institute for Applied Systems Analysis, and Wiley, Toronto, 1978).
- Pressey, R. L. in *Ecology for Everyone: Communicating Ecology to Scientists, the Public and the Politicians* (eds Wills, R. & Hobbs, R.) 73–87 (Surrey Beatty, Sydney, 1998).
- Cowling, R. M., Pressey, R. L., Lombard, A. T., Desmet, P. G. & Ellis, A. G. From representation to persistence: requirements for a sustainable reserve system in the species-rich Mediterranean-climate deserts of southern Africa. *Div. Distrib.* **5**, 51–71 (1999).
- Davis, F. W., Stoms, D. M. & Andelman, S. Systematic reserve selection in the USA: an example from the Columbia Plateau ecoregion. *Parks* **9**, 31–41 (1999).
- Nix, H. A. *et al.* *The BioRap Toolbox: A National Study of Biodiversity Assessment and Planning for Papua New Guinea. Consultancy Report to World Bank* (CSIRO Publishing, Melbourne, 2000).
- Noss, R. F., Strittholt, J. R., Vance-Borland, K., Carroll, C. & Frost, P. A conservation plan for the Klamath-Siskiyou ecoregion. *Nat. Areas J.* **19**, 392–411 (1999).
- Dobson, A. P., Rodriguez, J. P., Roberts, W. M. & Wilcove, D. S. Geographic distribution of endangered species in the United States. *Science* **275**, 550–553 (1997).
- Aiken, S. R. Peninsular Malaysia's protected areas' coverage, 1903–92: creation, rescission, excision, and intrusion. *Environ. Conserv.* **21**, 49–56 (1994).
- Pressey, R. L. *et al.* How well protected are the forests of north-eastern New South Wales?—Analyses of forest environments in relation to tenure, formal protection measures and vulnerability to clearing. *For. Ecol. Mgmt* **85**, 311–333 (1996).
- Ranta, P., Blom, T., Niemela, J., Joensuu, E. & Siitonen, M. The fragmented Atlantic rain forest of Brazil: size, shape and distribution of forest fragments. *Biodiv. Conserv.* **7**, 385–403 (1998).
- Sarkar, S. Wilderness preservation and biodiversity conservation—keeping divergent goals distinct. *BioScience* **49**, 405–412 (1999).
- Anon. *National Forest Policy Statement: a New Focus for Australia's Forests* (Australian Government Publishing Service, Canberra, 1992).
- Noss, R. F., O'Connell, M. A. & Murphy, D. D. *The Science of Conservation Planning: Habitat Conservation under the Endangered Species Act* (Island, Washington, 1997).
- Pressey, R. L. & Logan, V. S. in *Conservation Outside Nature Reserves* (eds Hale, P. & Lamb, D.) 407–418 (Univ. Queensland Press, Brisbane, 1997).
- Hansen, A. J., Garman, S. L., Marks, B. & Urban, D. L. An approach for managing vertebrate diversity across multiple-use landscapes. *Ecol. Applic.* **3**, 481–496 (1993).
- Prendergast, J. R. *et al.* Rare species, the coincidence of diversity hotspots and conservation strategies. *Nature* **365**, 335–337 (1993).
- van Jaarsveld, A. S. *et al.* Biodiversity assessment and conservation strategies. *Science* **279**, 2106 (1998).
- Howard, P. C. *et al.* Complementarity and the use of indicator groups for reserve selection in Uganda. *Nature* **394**, 472–475 (1998).
- McKenzie, N. L., Belbin, L., Margules, C. R. & Keighery, J. G. Selecting representative reserve systems in remote areas: a case study in the Nullarbor region, Australia. *Biol. Conserv.* **50**, 239 (1989).
- Goldstein, P. Z. Functional ecosystems and biodiversity buzzwords. *Conserv. Biol.* **13**, 247–255 (1999).
- Nix, H. A. in *Evolution of the Flora and Fauna of Australia* (eds Baker, W. R. & Greenslade, P. J. M.) 47–66 (Peacock, Adelaide, 1982).
- Nix, H. A. in *Atlas of Elapid Snakes of Australia* (ed. Longmore, R.) 4–14 (Australian Government Publishing Service, Canberra, 1986).
- Austin, M. P. Continuum concept, ordination methods and niche theory. *Annu. Rev. Ecol. Syst.* **16**, 39–61 (1985).
- Austin, M. P., Nicholls, A. O. & Margules, C. R. Measurement of the qualitative realised niche: environmental niches of five *Eucalyptus* species. *Ecol. Monogr.* **60**, 161–177 (1990).
- Wessels, K. J., Freitag, S. & van Jaarsveld, A. S. The use of land facets as biodiversity surrogates during reserve selection at a local scale. *Biol. Conserv.* **89**, 21–38 (1999).
- Richards, B. N. *et al.* *Biological Conservation of the South-East Forests* (Australian Government Publishing Service, Canberra, 1990).
- Ferrier, S. & Watson, G. *An Evaluation of the Effectiveness of Environmental Surrogates and Modelling Techniques in Predicting the Distribution of Biological Diversity* (Environment Australia, Canberra, 1997).
- Margules, C. R. & Austin, M. P. Biological models for monitoring species decline: the construction and use of data bases. *Phil. Trans. R. Soc. Lond. B* **344**, 69–75 (1994).
- Nelson, B. W., Ferreira, C. A. C., da Silva, M. F. & Kawasaki, M. L. Endemism centres, refugia and botanical collection intensity in Brazilian Amazonia. *Nature* **345**, 714–716 (1990).
- Hutchinson, M. F. *et al.* *BioRap Volume 2. Spatial Modelling Tools* (<http://cres.anu.edu/biorap/tools.html>) (The Australian BioRap Consortium, Canberra, 1996).
- Austin, M. P. & Meyers, J. A. Current approaches to modelling the environmental niche of Eucalypts: implications for management of forest biodiversity. *For. Ecol. Mgmt* **85**, 95–106 (1996).
- Deadman, P. J. & Gimblett, H. R. Applying neural networks to vegetation management plan development. *AI Applic.* **11**, 107 (1997).
- Jennings, M. D. Gap analysis: concepts, methods, and recent results. *Landscape Ecol.* **15**, 5–20 (2000).
- Pressey, R. L. & Logan, V. S. Size of selection units for future reserves and its influence on actual vs. targeted representation of features: a case study in western New South Wales. *Biol. Conserv.* **85**, 305–319 (1998).
- Soulé, M. E. & Sanjayan, M. A. Conservation targets: do they help? *Science* **279**, 2060 (1998).
- Commonwealth of Australia. *Nationally Agreed Criteria for the Establishment of a Comprehensive, Adequate and Representative Reserve System for Forests in Australia* (Australian Government Publishing Service, Canberra, 1997).
- Margules, C. R. & Nicholls, A. O. in *Nature Conservation: The Role of Remnants of Native Vegetation* (eds Saunders, D. A., Arnold, G. W., Burbidge, A. A. & Hopkins, A. J. M.) 89–102 (Surrey Beatty, Sydney, 1987).
- Balmford, A., Mace, G. M. & Ginsberg, J. A. in *Conservation in a Changing World* (eds Mace, G. M., Balmford, A. & Ginsberg, J. A.) 1–28 (Cambridge Univ. Press, Cambridge, 1998).
- MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography* (Princeton, New Jersey, 1967).
- Diamond, J. M. The island dilemma: lessons of modern biogeographic studies for the design of natural reserves. *Biol. Conserv.* **7**, 129 (1975).
- Wilson, E. O. & Willis, E. O. in *Ecology and Evolution of Communities* (eds Cody, M. L. & Diamond, J. M.) 522–534 (Belknap, Cambridge, MA, 1975).
- Higgs, A. J. Island biogeography and nature reserve design. *J. Biogeogr.* **8**, 117–124 (1981).
- Sauer, J. D. Oceanic islands and biogeographical theory. *Geogr. Rev.* **59**, 585 (1969).
- Davies, K. F., Margules, C. R. & Lawrence, J. F. Which traits of species predict population declines in experimental forest fragments? *Ecology* **81** (in the press).
- Levins, R. Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bull. Entomol. Soc. Am.* **15**, 237–240 (1969).
- Holt, R. D. & Gaines, M. S. in *Patch Dynamics* (eds Levin, S. A., Powell, T. M. & Steele, J. H.) 260–276 (Springer, Berlin, 1993).
- Gilpin, M. E. in *Viable Populations for Conservation* (ed. Soulé, M. E.) 126–139 (Cambridge Univ. Press, New York, 1987).
- Bennett, A. F. *Linkages in the Landscape: the Role of Corridors and Connectivity in Wildlife Conservation* (IUCN, Gland, Switzerland, and Cambridge, 1998).
- Thomas, C. D. *et al.* in *Conservation in a Changing World* (eds Mace, G. M., Balmford, A. & Ginsberg, J. R.) 107–138 (Cambridge Univ. Press, Cambridge, 1998).
- Holt, R. D. in *Species Diversity in Ecological Communities* (eds Ricklefs, R. E. & Schluter, D.) 77–96 (Univ. Chicago Press, Chicago, 1993).
- Lindenmayer, D. B. & Possingham, H. P. *The Risk of Extinction: Ranking Management Options for Leadbeater's Possum* (Centre for Resource and Environmental Studies, Australian National University, Canberra, 1995).
- Pickett, S. T. A. & Thompson, J. N. Patch dynamics and the design of nature reserves. *Biol. Conserv.* **13**, 27–37 (1978).
- Lamberson, R. H., Noon, B. R., Voss, C. & McKelvey, R. Reserve design for territorial species: the effects of patch size and spacing on the viability of the Northern Spotted Owl. *Conserv. Biol.* **8**, 185–195 (1994).
- Burgman, M., Ferson, S. & Akçakaya, H. R. *Risk Assessment in Conservation Biology* (Chapman & Hall, New York, 1993).
- Lindenmayer, D. B., Burgman, M. A., Akçakaya, H. R., Lacy, R. C. & Possingham, H. P. A review of three models for metapopulation viability analysis—ALEX, RAMAS/Space and VORTEX. *Ecol. Model.* **82**, 161–174 (1995).
- Lambeck, R. J. Focal species: a multi-species umbrella for nature conservation. *Conserv. Biol.* **11**, 849–856 (1997).
- Dias, P. C. Sources and sinks in population biology. *Trends Ecol. Evol.* **11**, 326–330 (1999).
- Braithwaite, L. W., Binns, D. L. & Nowlan, R. D. The distribution of arboreal marsupials in relation to eucalypt forest types in the Eden (NSW) Woodchip Concession Area. *Aust. Wildl. Res.* **10**, 231–247 (1988).
- Saunders, D. A., Hobbs, R. J. & Margules, C. R. Biological consequences of ecosystem fragmentation: a review. *Conserv. Biol.* **5**, 18 (1991).
- Kapos, V. Effects of isolation on the water status of forest patches in the Brazilian Amazon. *J. Trop. Ecol.* **5**, 173 (1989).
- Didham, R. K., Hammond, P. M., Lawton, J. H., Eggleton, P. & Stork, N. E. Beetle species responses to tropical forest fragmentation. *Ecol. Monogr.* **68**, 295–323 (1998).
- Margules, C. R., Milkovits, G. A. & Smith, G. T. Contrasting effects of habitat fragmentation on the scorpion, *Cercophonium squama* and an amphipod. *Ecology* **75**, 2033–2042 (1994).
- Cochrane, M. A. *et al.* Positive feedbacks in the fire dynamic of closed canopy tropical forests. *Science* **284**, 1832–1835 (1999).
- Yates, C. J. & Hobbs, R. J. Temperate eucalypt woodlands: a review of their status, processes threatening their persistence and techniques for restoration. *Aust. J. Bot.* **45**, 949–973 (1997).
- Frankel, O. H. & Soulé, M. E. *Conservation and Evolution* (Cambridge Univ. Press, Cambridge, 1981).
- Rojas, M. The species problem and conservation: what are we protecting? *Conserv. Biol.* **6**, 170–178 (1992).
- Vane-Wright, R. L., Humphries, C. J. & Williams, P. H. What to protect?—Systematics and the agony of choice. *Biol. Conserv.* **55**, 235–254 (1991).
- Fields, J. Geographic patterns for relic and young species of birds in Africa and South America and implications for conservation priorities. *Biodiv. Conserv.* **3**, 207–226 (1994).
- Kiester, A. R. *et al.* Conservation prioritization using GAP data. *Conserv. Biol.* **10**, 1332–1342 (1996).
- Kirkpatrick, J. B. An iterative method for establishing priorities for the selection of nature reserves: an example from Tasmania. *Biol. Conserv.* **25**, 127–134 (1983).
- Stoms, D. M. GAP management status and regional indicators of threats to biodiversity. *Landscape Ecol.* **15**, 21–33 (2000).
- Noss, R. F. Assessing and monitoring forest biodiversity: a suggested framework and indicators. *For.*

- Ecol. Mgmt* **115**, 135–146 (1999).
84. Williams, P. H. in *Conservation in a Changing World* (eds Mace, G. M., Balmford, A. & Ginsberg, J. R.) 211–249 (Cambridge Univ. Press, Cambridge, 1998).
 85. Faith, D. P. Phylogenetic pattern and the quantification of organismal biodiversity. *Phil. Trans. R. Soc. Lond. B* **345**, 45–48 (1994).
 86. Colwell, R. K. & Coddington, J. A. Estimating terrestrial biodiversity through extrapolation. *Phil. Trans. R. Soc. Lond. B* **345**, 101–108 (1994).
 87. Faith, D. P. & Walker, P. A. Environmental diversity: on the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. *Biodiv. Conserv.* **5**, 399–415 (1996).
 88. Belbin, L. Environmental representativeness: regional partitioning and reserve selection. *Biol. Conserv.* **66**, 223–230 (1993).
 89. Faith, D. P. & Norris, R. Correlation of environmental variables with patterns of distribution and abundance of common and rare freshwater macroinvertebrates. *Biol. Conserv.* **50**, 77–89 (1989).
 90. Williams, P. H. *et al.* Comparison of richness hotspots, rarity hotspots and complementary areas for conserving biodiversity, using British birds. *Conserv. Biol.* **10**, 155–174 (1996).
 91. Ferrier, S., Pressey, R. L. & Barrett, T. W. A new predictor of the irreplaceability of areas for achieving a conservation goal, its application to real-world planning, and a research agenda for further refinement. *Biol. Conserv.* (in the press).
 92. Faith, D. P., Margules, C. R., Walker, P. A., Hutchinson, M. & Nix, H. A. in *Science for Pacific Posterity: Environments, Resources and Welfare of the Pacific People* (ed. Anon) 153 (Univ. of New South Wales Press, Sydney, 1999).
 93. Faith, D. P., Walker, P. A., Ive, J. R. & Belbin, L. in *Conserving Biological Diversity in Temperate Forest Ecosystems—Towards Sustainable Management* (ed. Anon) 74–75 (Centre for Resource and Environmental Studies, Australian National University, Canberra, 1994).
 94. Faith, D. P. & Walker, P. A. in *BioRap Volume 3. Tools for Assessing Biodiversity Priority Areas* (eds Faith, D. P. & Nicholls, A. O.) 63–74 (The Australian BioRap Consortium, Canberra, 1996).
 95. Faith, D. P. & Walker, P. A. Integrating conservation and development: effective trade-offs between biodiversity and cost in the selection of protected areas. *Biodiv. Conserv.* **5**, 417–429 (1996).
 96. Pressey, R. L. in *National Parks and Protected Areas: Selection, Delimitation and Management* (eds Pigram, J. J. & Sundell, R. C.) 337–357 (Univ. of New England, Centre for Water Policy Research, Armidale, 1997).
 97. Myers, N. Threatened biotas: hotspots in tropical forests. *The Environmentalist* **8**, 178–208 (1988).
 98. Mittermeier, R. A., Myers, N., Thomsen, J. B., da Fonseca, G. A. B. & Olivieri, S. Biodiversity hotspots and major tropical wilderness areas: approaches to setting conservation priorities. *Conserv. Biol.* **12**, 516–520 (1998).
 99. Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
 100. Dinerstein, E. & Wikramanayake, E. D. Beyond “hotspots”: how to prioritize investments to conserve biodiversity in the Indo-Pacific region. *Conserv. Biol.* **7**, 53–65 (1993).
 101. Balmford, A. & Long, A. Avian endemism and forest loss. *Nature* **372**, 623–624 (1994).
 102. Sisk, T. D., Launer, A. E., Switky, K. R. & Erlich, P. R. Identifying extinction threats: global analyses of the distribution of biodiversity and the expansion of the human enterprise. *BioScience* **44**, 592–604 (1994).
 103. Ricketts, T. H. *et al.* *Terrestrial Ecoregions of North America: A Conservation Assessment* (Island, Washington DC, 1999).
 104. Myers, N. *The Sinking Ark: A New Look at the Problem of Disappearing Species* (Pergamon, Oxford, 1979).
 105. James, A. N. Institutional constraints to protected area funding. *Parks* **9**, 15–26 (1999).
 106. Stolton, S. & Dudley, N. A preliminary survey of management status and threats in forest protected areas. *Parks* **9**, 27–33 (1999).
 107. Caughley, G. & Sinclair, A. R. E. *Wildlife Management and Ecology* (Blackwell Science, Cambridge, MA, 1994).
 108. Hockings, M. & Phillips, A. How well are we doing?—Some thoughts on the effectiveness of protected areas. *Parks* **9**, 5–14 (1999).
 109. Faith, D. P. & Walker, P. A. in *National Parks and Protected Areas: Selection, Delimitation and Management* (eds Pigram, J. J. & Sundell, R. C.) 297–314 (Univ. New England Press, Armidale, 1997).

Acknowledgements

D. Faith, M. Hutchinson and H. Nix permitted the use of the unpublished map in Fig. 2. Many people have contributed to the ideas expressed here including M. Austin, C. Humphries, S. Ferrier, N. Nicholls, S. Sarkar, R. Vane-Wright, P. Walker and P. Williams. Critical comments from A. Balmford, G. Harrington, R. Noss and D. Westcott improved a draft of the manuscript. Some of the ideas discussed here were developed while both authors held fellowships at the Wissenschaftskolleg zu Berlin. T. Barrett and M. Watts prepared Figs 5 and 6.

Center for Applied Biodiversity Science



Investing in Conservation Solutions

Earth's biodiversity is being destroyed at an unprecedented rate, with far-reaching and irreversible consequences for life on our planet. If current trends continue unchecked, we can expect the end of nature as we know it. The threat to humanity is enormous, since we rely on healthy ecosystems for clean air and water, soil conservation, watershed protection, and pollination of major food crops, to name a few of nature's vital benefits.

Fortunately, the world is waking up to the escalating crisis. The 1992 Convention on Biological Diversity, signed by 175 countries, reflects a global consensus of the importance of biodiversity in maintaining the planet's life-sustaining systems. Yet, traditional reactive approaches will not suffice if we are to successfully confront the complex conservation challenges of this dawning millennium. All too often, conservationists, scientists, and decision-makers face major threats to biodiversity only after potentially manageable situations have solidified into intractable losses. The Center for Applied Biodiversity Science (CABS), a division of Conservation International, is working to change this equation by anticipating harmful scenarios before such situations become unsolvable crises. By providing early warning systems for threats to and opportunities for biodiversity and its many components, CABS is striving to make conservation a more proactive endeavour.

In order to fulfill its objectives, CABS is currently working with numerous universities, research centers, multilateral and

government agencies, non-governmental organizations, corporations, foundations, and other institutions. Each year, over 20 senior-level scientists from a range of backgrounds become CABS Fellows who put their expertise to work helping us identify threats to, and opportunities for, the conservation of biodiversity.

Research at the Center ranges from modeling tropical landscape dynamics to the economics of biodiversity protection. Working with its partners, CABS field-tests promising solutions in priority ecosystems that are distributed across the 25 countries where Conservation International has on-the-ground presence. The Center also explores trends and opportunities by holding high-level conferences and seminars. These meetings bring together scientists, decision-makers, business leaders, and the media. Hosted by Gordon Moore and Edward O. Wilson, the next major biodiversity conference will take place at the California Institute of Technology later this year.

Finally, an integral part of the Center's mission is to stimulate and promote the dissemination of high quality information and analyses pertaining to global conservation. We are proud, therefore, to be associated with this very timely supplement of *Nature*, which will prove an invaluable state-of-the-art reference in the applied biodiversity sciences.

Peter Seligmann

Chairman and CEO
Conservation International

Gustavo A. B. da Fonseca

Executive Director
Center for Applied Biodiversity Science

Integrating Science and Conservation

Center for Applied Biodiversity Science

BELOW RAP scientists, Dr. Marcos Callisto (Brazil) and Dr. LeeAnne Alonso (US) sample zooplankton in the Brazilian Pantanal, the world's largest wetland. (SEPTEMBER 1998)

The Center for Applied Biodiversity Science (CABS) is an institute designed to acquire knowledge most relevant to the assessment of endangered ecosystems and the unique species they contain, and to advise the conservation community on priorities this knowledge implies. Its mission also includes developing methods by which the goals of conservation can be more effectively met. CABS has an in-house function: to provide day-to-day scientific guidance for the programs of its parental organization, Conservation International, which is unusually practical and field-based in its orientation. It also aims to serve the broader global conservation community as an easily accessible research center and clearing house of information.

CABS researchers monitor and collate the now enormous and growing mass of information on biodiversity built by other scientists around the world. In addition, it draws on the original information provided by Conservation International's network of field stations, and still further on the databases from its Rapid Assessment Programs (RAPs). RAPs are surveys of vertebrates, plants, and other taxonomically well-known groups in habitats suspected of harboring significant numbers of unique species. When this initial sweep of sampling suggests that the locality in addition constitutes such a "hotspot" of biotic

endangerment, further studies and conservation activity can be quickly initiated. One of the key roles of CABS is thus to keep an eye on the natural world in order to provide an early warning system. In this it is true to the principle that conservation practice, like medicine, is most effective when correct diagnoses are made early.

The capacity of CABS and other organizations to master such a planet-wide set of problems is substantially enhanced by the ongoing revolution in information technology. It will soon be possible for researchers in the field, while conducting rapid assessment programs in even the most remote locations to verify taxonomic identifications by downloading illustrated monographs of plants, birds, and other organisms. When monographs are not available, the field biologist can acquire digitized images of type specimens kept in museums scattered around the world. Experts elsewhere can be consulted instantly by the same means, information can be exchanged, and advice can be sought. With the global positioning system the location of samples can be pinpointed to within a few tens of meters. And with key electronic journals online, other information on the ecology, geography, and economics of the region can be sifted and included, without delay. Much of the basic research for conservation practice will thus be speeded up, very likely by one or two orders of magnitude, as a great deal of the primary research, analysis, and reporting are accomplished right at camp sites.

By necessity, the role of CABS also extends beyond the natural sciences into the social sciences. Conservation is not achieved by the creation of reserves alone. It requires the support of local people whose lives are improved by the security of the natural environment around them, and of political leaders who come to view conservation as a benefit to their constituencies and themselves. In short, conservation science must include research in the relevant aspects of economics and the other social sciences, fitted hand in glove to traditional biodiversity research.

Gordon Moore

Co-Founder & Chairman Emeritus
Intel Corporation

Edward O. Wilson

University Research Professor
Harvard University



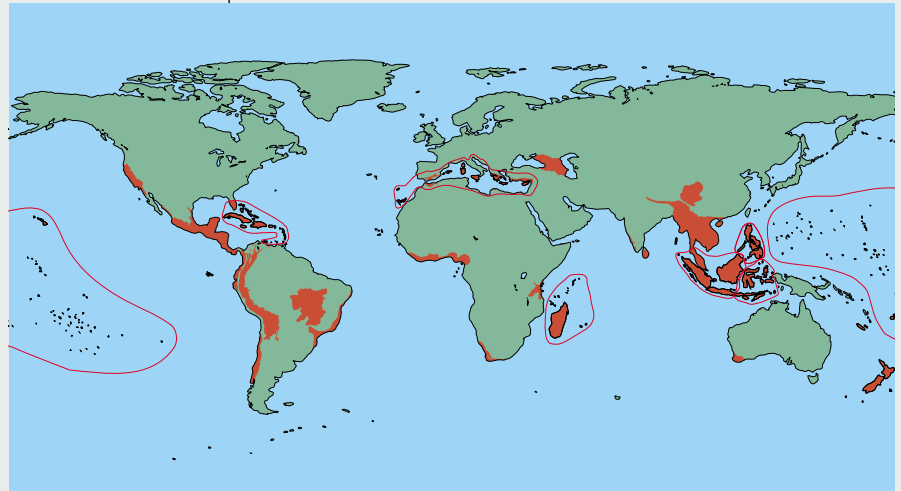
Conservation International and Biodiversity Conservation

Since its creation in 1987, the mission of Conservation International (CI) has been to conserve our planet's biodiversity and to demonstrate that human societies can live harmoniously with nature. This mission is both science-based and people-centered, recognizing that conservation must rely upon the soundest scientific underpinnings, while at the same time addressing the needs and aspirations of human communities.

The core element of CI's strategic approach is the hotspots concept. First developed by British ecologist Norman Myers, it focuses heavily on areas that are rich in biodiversity, and especially in endemic species, while being under severe threat. CI has just completed a reanalysis of the hotspots, a three and a half year study involving some 100 experts from around the world (see results in *Nature* 403:853-858, 2000). Twenty-five terrestrial hotspots were identified, including several not previously recognized (see map). These areas once covered about 12% of the land surface of the planet, but cumulatively have lost 88% of their original natural vegetation; they are now down to just 1.4% of the Earth's land area. Although this area is only about three times the size of Texas, the hotspots harbor, as endemics, fully 44% of all vascular plants and 35% of all tetrapods (i.e., mammals, birds, reptiles, and amphibians).

Paralleling the hotspots strategy is the major tropical wilderness areas approach. As with hotspots, these areas are also exceptionally rich in biodiversity; unlike the hotspots, they are still largely (>75%) intact, with low human population. Few and far between, these wilderness areas are found mainly in western and central Amazonia, the Guayana Shield region of northeastern Amazonia, the Congo Basin, and the island of New Guinea. About 60% of CI's programs focus on 13 of the 25 hotspots, while the remaining 40% are concentrated in the four wilderness areas.

CI has long been a leader in biodiversity conservation, having introduced innovative approaches and tools to the business of saving the full range of life on Earth. Included among its many firsts are the first-ever "Debt for Nature Swap," the user-friendly conservation software CI/SIG tailored to the needs of developing countries, and the Rapid Assessment Program (RAP) designed to



THE 25 GLOBAL BIODIVERSITY HOTSPOTS

quickly and effectively survey poorly known tropical ecosystems. CI has also been a pioneer in engaging the private sector in conservation, including partnerships with Ford Motor Company to promote conservation in Brazil, with Intel Corporation to bring leading edge information technology to the field, and with Starbucks to market environmentally-friendly coffee from hotspot ecosystems.

Major new efforts include the Healthy Communities Initiative (1997), which investigates linkages between biodiversity and human health, and the Tropical Wilderness Protection Fund (1999), which focuses on setting aside large blocks of intact tropical forest. Finally, CI's Center for Applied Biodiversity Science (CABS), established in 1999 and sponsor of this landmark supplement, is providing scientific leadership to the conservation community at large. In general, CI places great emphasis on partnerships with all sectors of society, since we believe that they are vital to making biodiversity conservation a reality.

CI is headquartered in Washington, D.C. and has offices in London, Tokyo, Amsterdam, Cape Town, and most of the 25 tropical countries in which it works in the field. More than 75% of the organiza-

tion's 600 staff are based in the field and represent more than 30 different nationalities. Its budget in fiscal year 2000 is \$33.6 million.

The organization is now positioning itself to set new standards for biodiversity conservation, among them a "moral high ground" of "zero deforestation – zero further species loss" for the hotspots. The first project under this banner is focusing on the Brazilian Atlantic Forest, through a major partnership with the Fundacao SOS Mata Atlantica, the largest Brazilian NGO. To further these and other major objectives, CI is now in the middle of a five-year, \$200 million capital campaign to raise the resources necessary for its ambitious conservation program in the new millennium.

Most important of all, CI has a positive, optimistic, can-do philosophy to achieving its goals. We believe that flexibility, agility, and the ability to adapt and create new and innovative solutions can take the conservation movement a long way towards successfully meeting the enormous challenges of the first two decades of the new millennium.

Russell A. Mittermeier

President
Conservation International



Making the most of DNA

Aids to DNA synthesis, repair, amplification and the rest.

PowerScript

Clontech

www.clontech.com

Reverse transcriptase specially prepared for cDNA synthesis

PowerScript is a point mutant of Moloney murine leukaemia virus (MMLV) reverse transcriptase (RT). This enzyme lacks RNase H activity but retains wild-type polymerase activity, so can synthesize longer cDNA fragments than wild-type MMLV RT. Clontech say that PowerScript achieves exceptional performance in long-distance cDNA synthesis—in-house experiments have shown that cDNA fragments of greater than 3 kb are generated. PowerScript preparations exhibit virtually no detectable RNase or DNase activity.

Reader Service No. 100

Phase lock gel

Eppendorf

www.eppendorfsci.com

Optimize DNA and RNA extraction

0.5 ml Phase Lock Gel is an additive which substantially accelerates and simplifies phenol or chloroform extractions of nucleic acids. The manufacturers claim that nucleic-acid yields can be increased by 30%, with the biggest benefit in the phenolization of small volumes. DNA or RNA purity is maximized by the elimination of contamination with interphase material. The gel forms a solid barrier between the aqueous and the organic phases during centrifugation, so the organic phase is entirely

enclosed by a gel barrier. The aqueous phase can be withdrawn and recovered with a pipette. 0.5 ml Phase lock gel is especially well suited for use with restriction or modification enzymes. Reactions can be carried out in the presence of PLG, stopped by adding phenol or chloroform and immediately extracted.

Reader Service No. 101

Reagents

Cruachem, Ltd.

www.cruachem.com

DNA and RNA synthesis reagents

Cruachem's DNA-synthesis reagents are compatible with Cruachem PS250 and all ABI DNA synthesizers. Cyanoethylphosphoramidites available include T-CE, dG-CE, dA-CE and dC-CE derivatives. New to the range is acetyl dC-CE phosphoramidite. Both purity and coupling yields are guaranteed at greater than 99%. New specialty synthesis reagents include reverse phosphoramidites, H-phosphonate monomers and reporter molecules.

Reader Service No. 102

StratoSpheres

Polymer Laboratories

www.polymerlabs.com

Aiming high?

The StratoSpheres range is produced using copolymerization to generate high-quality resin supports for solid-phase synthesis. The level of loading and homogeneity is tightly controlled directly from the base chloromethyl styrene resin copolymer. These resins are intended to eliminate the side reactions which can introduce anomalous functional groups or additional crosslinking. StratoSpheres resins are sized to produce a uniform particle size distribution for free draining and elimination of blocked filters.

Reader Service No. 103

North2South

Pierce Chemical

www.piercenet.com

A robust chemiluminescent nucleic-acid blotting kit

The North2South Chemiluminescent Hybridization and Detection Kit includes optimized hybridization and stringency wash buffers aimed at reducing background problems and inconsistent results associated with chemiluminescent blotting. Post-hybridization time is about 60 minutes, and immediate signal generation and brief film exposure make the North2South system "hours faster than DIG and days faster than ^{32}P ", say Pierce. Sensitivity is equivalent to that obtained with ^{32}P , but North2South is nonradioactive.

Reader Service No. 104

DNAzol ES...

Molecular Research Center www.mrcgene.com

... where ES stands for Extra Strength

A novel guanidine-detergent lysing mixture is included in DNAzol ES to maximize DNA extraction, and help keep contamination to a minimum during plant genomic DNA isolation. DNAzol ES is a complete and ready-to-use reagent with a protocol that includes optional steps appropriate for more difficult plant tissues. For 1 g of plant tissue, just 3 ml of DNAzol ES is needed for DNA extraction.

Reader Service No. 104

Gene Detect kits

GeneExpression Systems

Quality checker for cDNA

Human Gene Detect and Mouse Gene Detect Kits are recommended as a check on the quality of cDNA synthesis, and cDNA library construction. The kits simplify the detection of tube-to-tube variation in RT-PCR, measuring the purity of RNA and DNA during nucleic-acid isolation procedures. Comparative gene-expression profiles of any interesting target gene can be determined in a single tube or multiple tubes by normalizing with the housekeeping gene ExpressionMers supplied in these kits. Each kit consists of four pairs of ExpressionMers especially for highly abundant, moderately abundant and low abundant genes in order to measure the steady-state expression levels. A single kit provides sufficient reagents to perform a total of 200 reactions.

Reader Service No. 105



From Eppendorf: it's a phase they go through.



Vial bodies: Gene Detect, for checking cDNA.

new on the market

Latitude

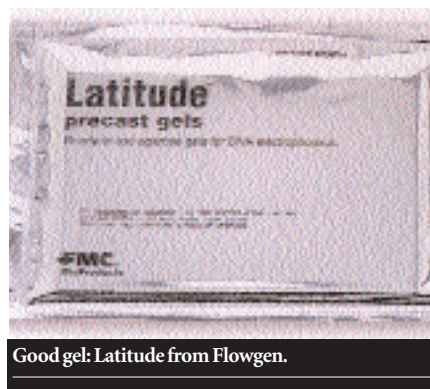
Flowgen

www.flowgen.co.uk

Rapid agarose gel separation of DNA

Latitude precast horizontal midgels are intended to provide researchers with a fast and reliable separation of DNA of all sizes. Compared to reliant gels, Latitude gels are larger with more sample wells, compatible with multichannel pipette loading. The additional space also enables finer resolution of DNA. Latitude gels are cost effective, as there is no preparation time. They are simple to use and results are ready in less than 30 minutes. Latitude gels are manufactured with molecular biology grade agaroses in a variety of concentrations and buffers.

Reader Service No. 107



second-strand cDNA synthesis of tough or rare messages up to 6.5 kb.

Reader Service No. 109

Antibodies

Novus Biologicals www.novus-biologicals.com

Tools for DNA repair and hypoxia research

Novus Biologicals, Inc. has introduced a range of research products especially for the study of DNA repair mechanisms and hypoxia. These tools include antibodies, cDNA and control peptides. Recent releases include antibodies to hOgg1 (8-oxoguanine DNA glycosylase), HIF-2 alpha (EPAS1), and APE/Ref-1 Monoclonal.

Reader Service No. 108

RobusT RT-PCR

Finnzymes Oy

www.finnzymes.fi

High fidelity one-tube RT-PCR system with Dynazyme EXT

Developed by Finnzymes Oy and distributed in the United States by MJ Research, Inc., the RobusT RT-PCR kit comprises (1) AMV Reverse Transcriptase (AMV-RT), to perform first-strand cDNA reactions at up to 60 °C to melt out any secondary structure. And (2), Dynazyme EXT DNA polymerase, a high performance *T. brockianus*-based enzyme for high fidelity, highly G + C-rich sequences and long templates. The kit can perform accurate

RNase H Minus

Promega Corporation

www.promega.com

Go into reverse

With cDNA synthesis with long mRNA templates in mind, Promega have added M-MLV Reverse Transcriptase, RNase H Minus, to their range. Other applications for which it is suitable include RT-PCR. M-MLV RT (H-) is available in catalogue package sizes as well as in bulk quantities.

Reader Service No. 110

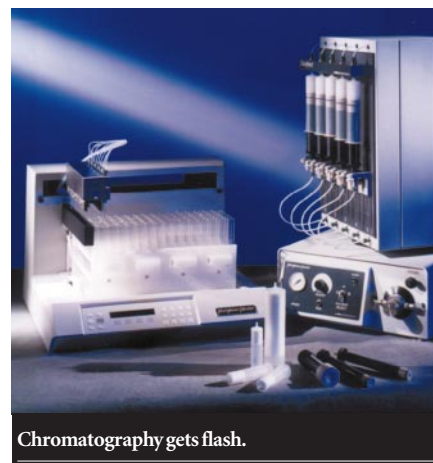
FlashMaster Parallel

Jones Chromatography

www.jones-chrom.co.uk

Fast throughput in parallel flash chromatography

This recent addition to the FlashMaster chromatography range is designed to speed up throughput in parallel organic synthesis purification. The system can be coupled to commercial parallel organic synthesizers for online synthesis purification. Features include parallel purification of ten samples, a small footprint, gradient formation capability, and a disposable column system to



eliminate downtime caused by emptying and repacking of glass columns.

Reader Service No. 111

NucleoScan 2000

NucleoTech Corporation

www.nucleotech.com

Real-time gel electrophoresis

The NucleoScan 2000 is a DNA fragment analyser based on an ultra-thin gel system. It can be used to perform microsatellite analysis in 30 minutes, run SSCPs in an hour, and mutation screenings in less than 40 minutes. NucleoTech also claim that NucleoScan 2000 can even perform DNA sequencing of up to 300 bases. Peltier-controlled temperature regulation contributes to fast run times. A range of fluorescent labels can be used, including ethidium bromide. NucleoTech's GelExpert software can be used with the 2000 system.

Reader Service No. 112

ATM antibody

Upstate Biotechnology

www.upstatebiotech.com

Antibodies to a protein kinase essential for p53 phosphorylation following DNA damage

Mutations in the ATM (ataxia-telangiectasia mutated) gene cause the rare human disease ataxia telangiectasia. Inherited mutations in this gene lead to a syndrome characterized by neuronal degeneration, oculocutaneous telangiectasia, immune dysfunction, cancer predisposition and premature ageing. Recent work has implicated signalling cascades downstream of ATM which guard genomic integrity after DNA damage. Following double-strand breaks in DNA, p53 becomes induced and is phosphorylated at Ser 15. This process has an absolute requirement for ATM. Upstate Biotechnology now announce the availability of antibody directed against ATM.

Reader Service No. 113

These notes are compiled in the Nature office from information provided by the manufacturers. For more details, fill in the reader service card.

ADVERTISEMENTS

Look no further

for the best value in

Oligo Synthesis

Competitive pricing
Routine 24 hour dispatch • Every Oligo QC verified by PAGE
Rigorous batch QC by PCR and Mass Spec
Web ordering and tracking • Local Freephone numbers
A great deal in any language!!

SIGMA GENOSYS
(+44) (0) 1223 839000 • Fax (+44) (0) 1223 839200
email: info@sigma-genosys.co.uk • www.sigma-genosys.co.uk

BACs • YACs • PACs

cDNAs

Research Genetics provides clones and colony membranes for a number of human, mouse and rat libraries as well as a custom screening service for the libraries.

Research Genetics
U.S. or Canada 800-533-4363
U.K. 0-800-89-1393
FAX 256-536-9016

Check out our homepage at
<http://www.resgen.com>

READER ENQUIRY NO. 33

READER ENQUIRY NO. 40
© 2000 Macmillan Magazines Ltd

ECONOMIES OF SCALE

The whys and wherefores of drug company mergers

p257

PROPOSALS OF MARRIAGE

How two companies are positioning themselves for the future

p258

MONEY MATTERS

Who earns what in the pharmaceutical business

p258

Are mega-mergers good medicine for the pharmaceutical industry?

Joining forces creates quick savings. But can vast organizations promote the kind of innovative research they need? Diane Gershon finds out.

The merger of Glaxo Wellcome and SmithKline Beecham and the takeover of Warner-Lambert by Pfizer, when finalized, will create industrial giants. At the same time they raise interesting questions. Is bigger always better? And are these mega-mergers, which are now wending their way through the process of regulatory and shareholder approval, focused on the short or the long term?

According to PricewaterhouseCoopers' 1998 report *Pharma 2005: An industrial revolution in R&D*, the pharmaceutical industry "faces a challenge of absolutely unprecedented scale". Research and development (R&D) costs are escalating, prescription drug sales are slowing down, product life cycles are getting shorter and the demographic profile of the industry's customer base is shifting.

At the same time, scientific and technological advances are poised to revolutionize the drug discovery process. Technologies such as combinatorial chemistry and high-throughput screening (HTS) are producing more new drug leads, all of which must compete for resources. Part of the challenge will be to improve the quality of the leads generated and focus on those with the highest revenue potential.

The top 20 companies saw R&D costs more than double over the seven years to 1998, according to the report, while growth in prescription sales has slowed. Bringing a drug to market is now estimated to cost \$500 million. In the 1980s, the industry spent 10 per cent of sales revenues on R&D in a market growing at about 11 per cent a year. In 1997, sales at 35 respondent companies were up just 6 per cent on the previous year, with an average 17.5 per cent of income from sales spent on R&D. Though the leaders fared slightly better, with sales up an average of 14 per cent, R&D still accounted for 16.8 per cent of sales revenues.

"The problem is, research is giving us a lot of compounds, but the process to develop them is hit and miss, very costly and time-consuming and the results are not quite keeping up with [investor] expectations,"



Welcome news: Glaxo Wellcome's R&D facility in Stevenage will survive the merger.

says Joe Palo of PricewaterhouseCoopers.

One option is to merge. In the short term, mergers can create value in terms of market position, franchise creation, cost-savings and efficiencies. The merger of two companies "gives you a one-time saving of about 15 per cent," says Palo. "You have a bigger war chest, so you can place more bets — put more molecules into development." It also offers a company more options on managing its portfolio and committing its sales force, he says.

"There is an argument [for scale] at two ends of the spectrum: the marketing end and the discovery end," says Stewart Adkins of Lehman Brothers, London. Glaxo SmithKline plans to have a worldwide sales and marketing force of about 40,000, with some 7,200 sales representatives in the United States alone, where its operational headquarters will also be located.

Glaxo SmithKline's decision to increase its presence in the United States reflects an

important industry-wide trend. The United States offers the largest and most lucrative unified market for drug sales. Its lack of price controls allows greater profit margins. Moreover, the US population is increasing in number, in age and in lifestyle expectations, stimulating the demand for new kinds of medicines. And in the US 'managed care' system, drugs are often promoted as a cost-effective alternative to other treatments.

IMS Health, a global healthcare information company that tracks drug sales through retail pharmacies in key global markets, notes that the United States continues to lead the growth in sales. In the 12 months to February 2000, sales there and in Canada grew by 15 per cent and were valued at \$92.3 billion (of which the United States accounted for \$87.4 billion). Growth in Europe was just 8 per cent, with sales in the top five European countries (Germany, France, Italy, the United Kingdom and Spain) totalling \$53.7 billion. The Japanese retail drug market, valued at \$48.2 billion, grew 6 per cent.

In the United States, large companies are now beginning to play the 'megabrand game', says Adkins, committing up to \$1 billion for the first couple of years of a product launch in an effort to gain a large chunk of market share.

Though mergers can provide cost-savings

New names from old

- Astra AB and Zeneca Group plc (AstraZeneca)
- Hoechst AG and Rhône-Poulenc SA (Aventis)
- Pharmacia & Upjohn and Monsanto (Pharmacia)
- Pfizer and Warner-Lambert (Pfizer)
- Glaxo Wellcome and SmithKline Beecham (Glaxo SmithKline)

in the short term, the essence of long-term survival in pharmaceuticals is innovation", says Charles Wheeler of the Boston Consulting Group.

A certain size is required for investment in platform technologies such as bioinformatics, combinatorial chemistry, HTS and pharmacogenetics, which can generally be shared across different programmatic research areas. The counter-argument, he says, would be that the days of hierarchical owning-it-all organizations are numbered — many innovative drugs are now coming from companies "that can be flexible and nimble and can move from relationship to relationship". He says some companies, such as Eli Lilly, maintain that research is a "critical-mass business, not a scale business", and that above critical mass, scale is not a factor.

The challenge, says Wheeler, will be creating an efficient research organization on a

large scale. Research scientists may find such big organizations unappealing, unless the companies set up structures that foster a creative culture such as that characteristic of the biotechnology industry. There are also alternative models that can be used, such as the recently formed Genomics Institute of the Novartis Research Foundation in La Jolla, California.

The merger of Glaxo Wellcome and SmithKline Beecham will create a company with an estimated global market share of 7.3 per cent. Nevertheless, this is still a fragmented industry, with ample room for further consolidation. Adkins says that he would be surprised if there wasn't at least one more deal in the next 12 months. The PricewaterhouseCoopers report went further, in predicting significant industry restructuring, with perhaps as few as 13 top companies remaining by 2005. ■

Partners resolve their differences and unite at the second attempt

Glaxo Wellcome and SmithKline Beecham's announcement, in February, of their intention to join forces was met with an air of inevitability: the companies had been there before. Talks in 1998 broke down reportedly because of disagreements between top executives.

On the face of it, the reasons for this merger and for Glaxo's takeover of Wellcome in 1995 seem very different. Glaxo enjoyed significant growth during the 1980s, thanks to its anti-ulcer drug Zantac, which, at its peak, represented almost half the company's sales. But the expiry of Zantac's patent was looming. Glaxo's answer was to acquire UK-based Wellcome plc, a company one-third its

size, with a similarly strong R&D tradition but with less of a commercial edge because of its charitable origins: it was almost 40 per cent owned by the Wellcome Trust.

By contrast, Glaxo Wellcome faces no major patent expiries over the next few years. The current merger "is fundamentally about R&D and really harnessing and having the [scale and] resources available to take advantage of all the new technologies," says Martin Sutton, head of Glaxo Wellcome's corporate communications in London. The company could have gone along without merging, "but it provided the opportunity to create what we think is going to be a terrific R&D powerhouse in the industry".

What are you worth to your employer?

A 1998 salary and employment status survey carried out by the American Association of Pharmaceutical Scientists (AAPS; <http://www.aaps.org>) of Arlington, Virginia, outlines the demographic and economic characteristics of the association's membership.

The survey is based on a questionnaire filled in by 1,538 of its members. The majority were men (73%) and the average age of respondents was 42. Most (84%) were employed in industry, 12% in academia and fewer than 2% in government jobs.

In 1998, average salaries were \$89,100 in industry (up 3% from the previous year), \$81,900 in academia (down 1%) and \$79,500 for those in government jobs (up 5%). Overall, the best paid had degrees in pharmacy administration (average salary of \$119,900), clinical pharmacology (\$117,300), clinical pharmacy (\$109,200) and pharmacology (\$106,000). At the lower end of the scale came those with degrees in chemistry (\$71,400), inorganic chemistry (\$70,900) and microbiology (\$70,700). The highest salaries in industry were paid for managerial positions, such as general management (\$127,000, down 6% on 1997 figures) and R&D management (\$117,500, up 3%), followed by clinical researchers (\$102,100, up 7%) and those in regulatory affairs (\$98,400, up 5%).

For people embarking on their career in 1998, the average starting salary was \$53,300, a drop of 13% from the previous year. **D.G.**

Billed as a merger of equals (although Glaxo Wellcome and SmithKline Beecham shareholders will hold 58.75 per cent and 41.25 per cent, respectively, of share capital), Glaxo SmithKline will have a combined market capitalization of \$189 billion, pharmaceutical sales of \$20.9 billion (based on 1998 pro forma information) and a current annual R&D budget of about \$4 billion.

The merger is expected to provide annual savings of \$1.7 billion after three years, of which a quarter will be re-invested in R&D. Job losses are inevitable, the companies say. However, unlike the Glaxo Wellcome merger, which closed Wellcome's site in Beckenham, the companies do not anticipate the closure of any major R&D sites. When the merger is complete, the company will move its operational headquarters from Britain to the United States, although London will remain as Glaxo SmithKline's corporate HQ. Based on combined 1998 sales data, Glaxo SmithKline would have derived about 45 per cent of its revenues from the United States and 33 per cent from Europe (of which the United Kingdom represents about 6 per cent).

The merger, contingent upon regulatory and shareholder approval, is expected to be completed by summer. ■

Diane Gershon is Assistant Editor, New Technology, at Nature Medicine.

Combined harvesters in the drug field

	Glaxo Wellcome 60,000 employees	SmithKline Beecham 47,000 employees
Top-selling drugs	Flixotide/Flovent: inhaled corticosteroid for asthma Imigran/Imitrex: migraine Zantac: anti-ulcer, reflux (sales now slipping)	Seroxat/Paxil: anti-depressant Augmentin: anti-infective
Strengths	Antiviral (HIV), respiratory and CNS therapeutic areas Affymax Research Institute: high-throughput drug synthesis and screening technologies	Newly launched diabetes drug Avandia Vaccines business Consumer healthcare and over-the-counter business Bioinformatics
Sales*	\$13.7 billion	\$13.6 billion
R&D spending*	\$2.06 billion	\$1.6 billion
* 1999 figures		